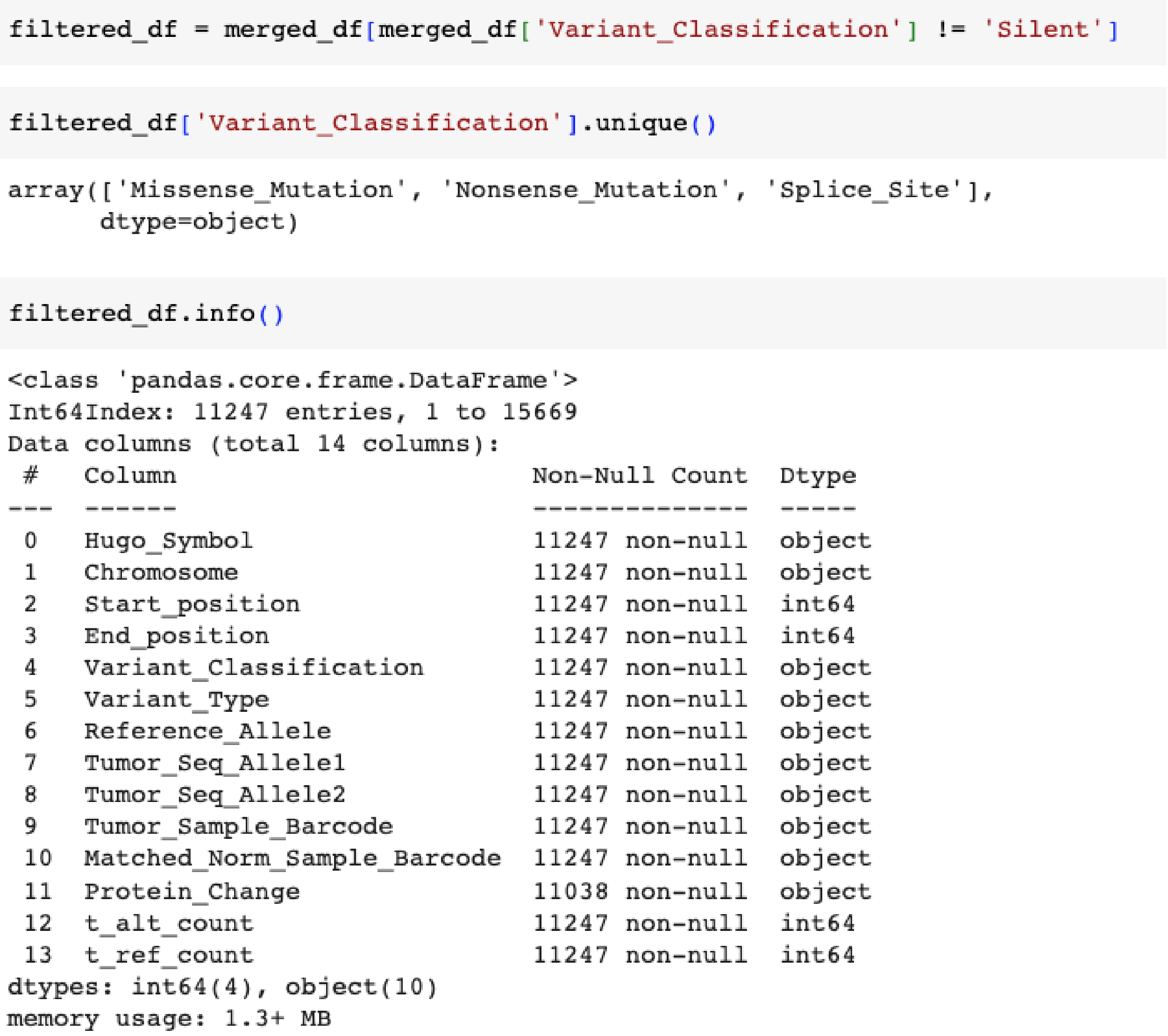
Report Summary:

I strongly recommend following the link to the python file attached and opening it in google-colab for an interactive experience and more step-wise detailed explanation below.

https://colab.research.google.com/drive/1HBX728ovsY7QYCb2cSDXU4p9k_pbCTse?usp=sharing#scrollTo=TQtHqbuND2IW

Overall question: Can you discover any mutations that are associated with treatment response?

1. **Subset for mutations that are not of the Variant Classification "Silent". For the purposes of this analysis, we will restrict ourselves to substitutions which result in changes to the produced protein ("nonsynonymous mutations").**



```
filtered_df = merged_df[merged_df['Variant_Classification'] != 'Silent']
```

```
filtered_df['Variant_Classification'].unique()
```

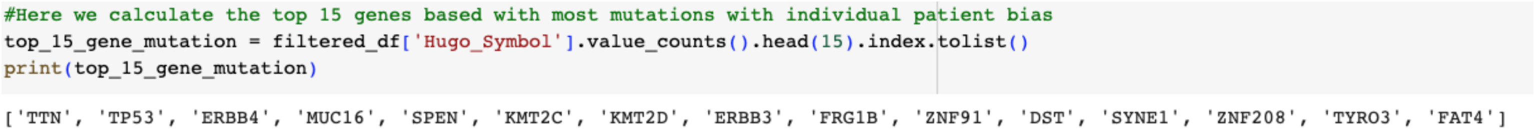
```
array(['Missense_Mutation', 'Nonsense_Mutation', 'Splice_Site'],
      dtype=object)
```

```
filtered_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 11247 entries, 1 to 15669
Data columns (total 14 columns):
 #   Column                   Non-Null Count  Dtype 
---  ------                   --------------  ----- 
 0   Hugo_Symbol              11247 non-null  object
 1   Chromosome               11247 non-null  object
 2   Start_position           11247 non-null  int64 
 3   End_position             11247 non-null  int64 
 4   Variant_Classification   11247 non-null  object
 5   Variant_Type             11247 non-null  object
 6   Reference_Allele         11247 non-null  object
 7   Tumor_Seq_Allele1        11247 non-null  object
 8   Tumor_Seq_Allele2        11247 non-null  object
 9   Tumor_Sample_Barcode     11247 non-null  object
 10  Matched_Norm_Sample_Barcode  11247 non-null  object
 11  Protein_Change           11038 non-null  object
 12  t_alt_count              11247 non-null  int64 
 13  t_ref_count              11247 non-null  int64 
dtypes: int64(4), object(10)
memory usage: 1.3+ MB
```

2. **Find the 15 most common mutations. Gene names are included in the column Hugo_Symbol and protein changes are stored in the column Protein_Change.**

['TTN', 'TP53', 'ERBB4', 'MUC16', 'SPEN', 'KMT2C', 'KMT2D', 'ERBB3', 'FRG1B', 'ZNF91', 'DST', 'SYNE1', 'ZNF208', 'TYRO3', 'FAT4']



```
#Here we calculate the top 15 genes based with most mutations with individual patient bias
top_15_gene_mutation = filtered_df['Hugo_Symbol'].value_counts().head(15).index.tolist()
print(top_15_gene_mutation)
```

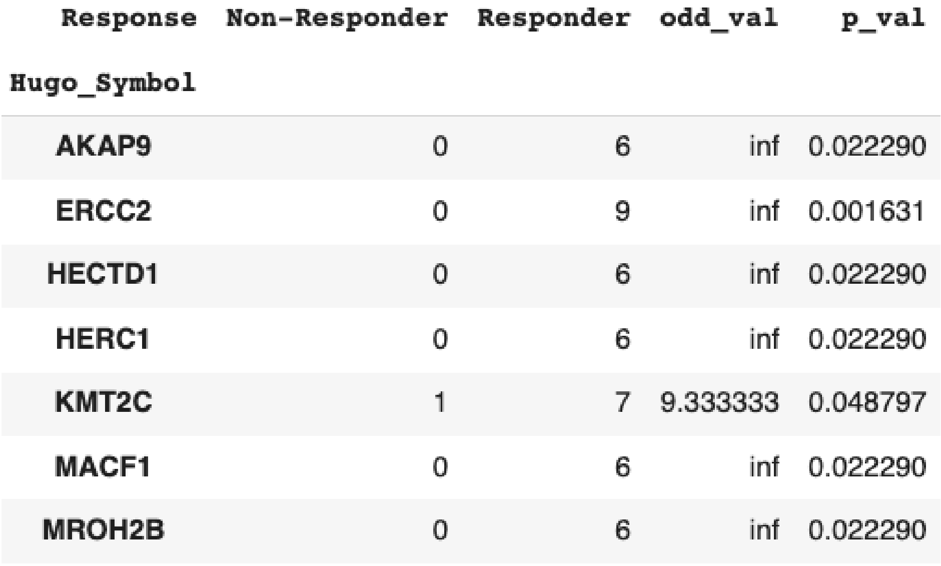
```
['TTN', 'TP53', 'ERBB4', 'MUC16', 'SPEN', 'KMT2C', 'KMT2D', 'ERBB3', 'FRG1B', 'ZNF91', 'DST', 'SYNE1', 'ZNF208', 'TYRO3', 'FAT4']
```

Above are top 15 gene mutated in most patients.

I also filtered top 15 protein mutations overall in the sample population.

3. **Perform a statistical test to explore if any mutated genes are enriched in patients who either responded or not. Response labels for individual patients are found in the file data/sample-information.tsv.**

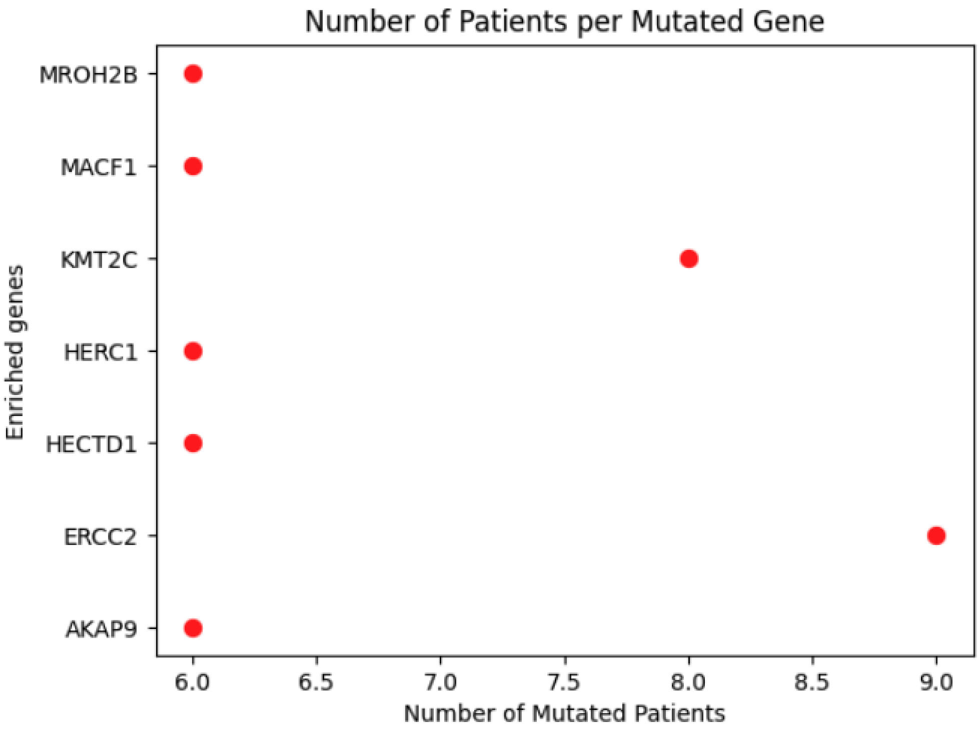
Statistically significant genes are listed below by Fischer's-exact test. Alpha = 0.05 due to low positives although we could use 0.025 and filter KMT2C out.



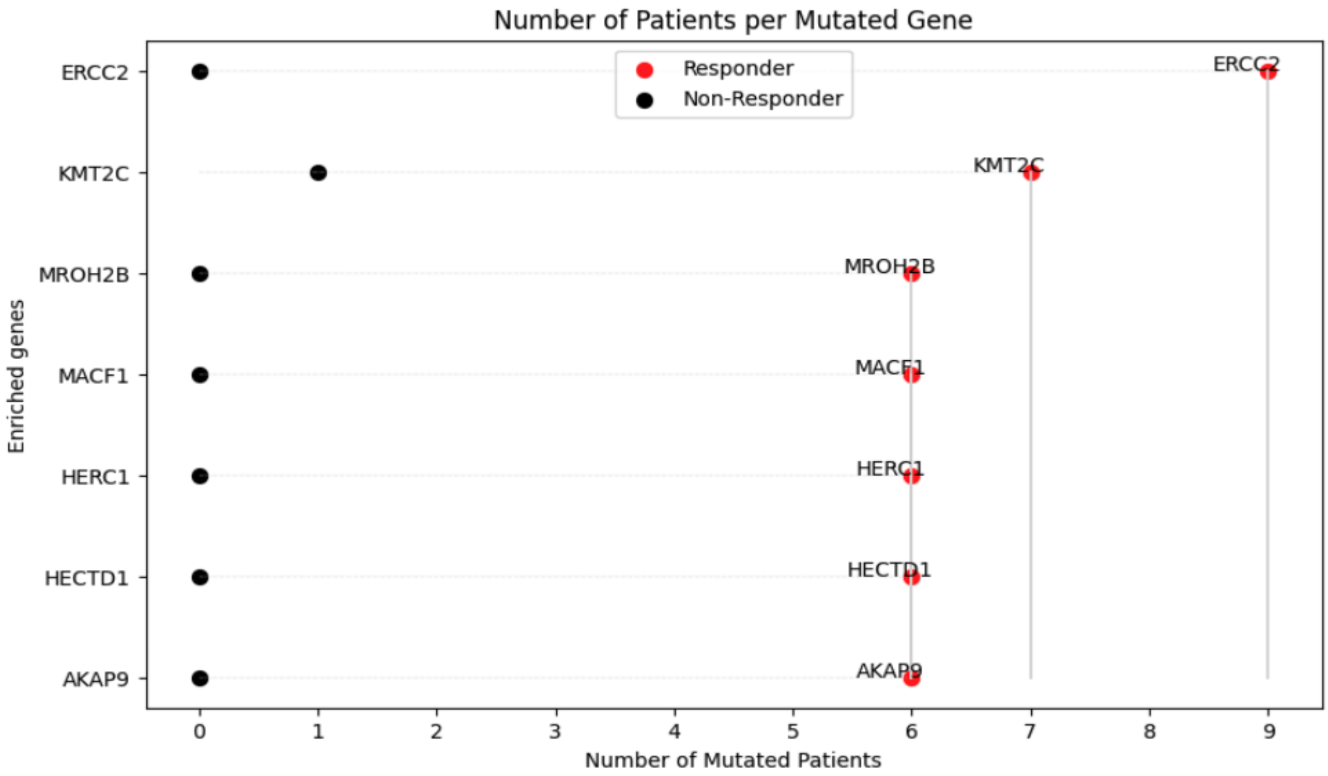
Response	Non-Responder	Responder	odd_val	p_val
Hugo_Symbol				
AKAP9	0	6	inf	0.022290
ERCC2	0	9	inf	0.001631
HECTD1	0	6	inf	0.022290
HERC1	0	6	inf	0.022290
KMT2C	1	7	9.333333	0.048797
MACF1	0	6	inf	0.022290
MROH2B	0	6	inf	0.022290

4. **Create a scatter plot of genes with the number of mutated patients on the x-axis and your results from question 4 on the y-axis. Can the figure in any way to improve readability? If so, recreate the plot using your suggestion(s).**

Original Plot



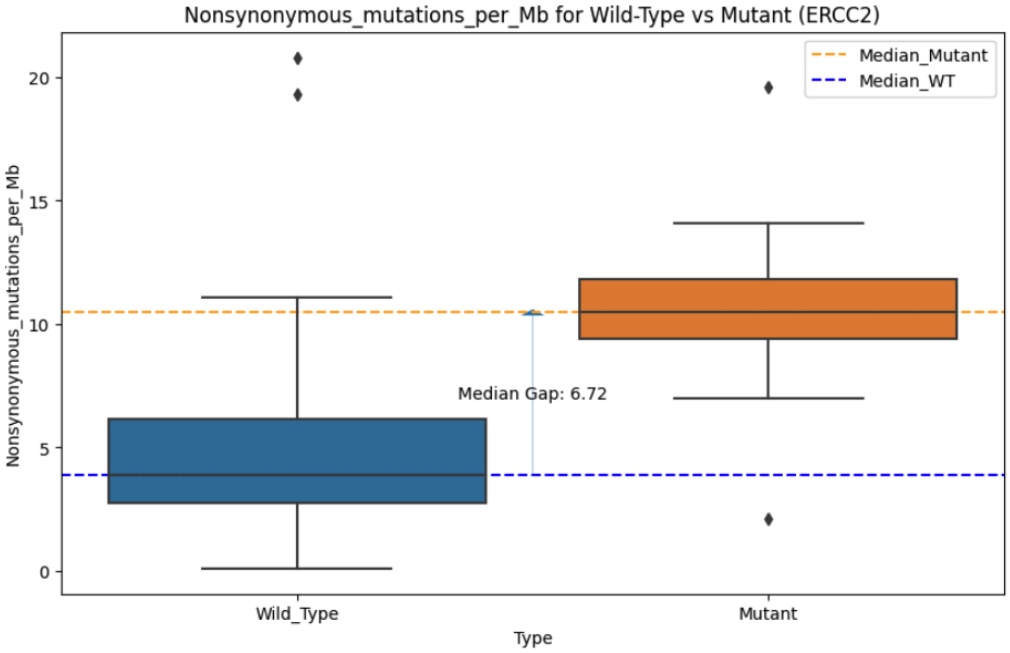
Recreated plot

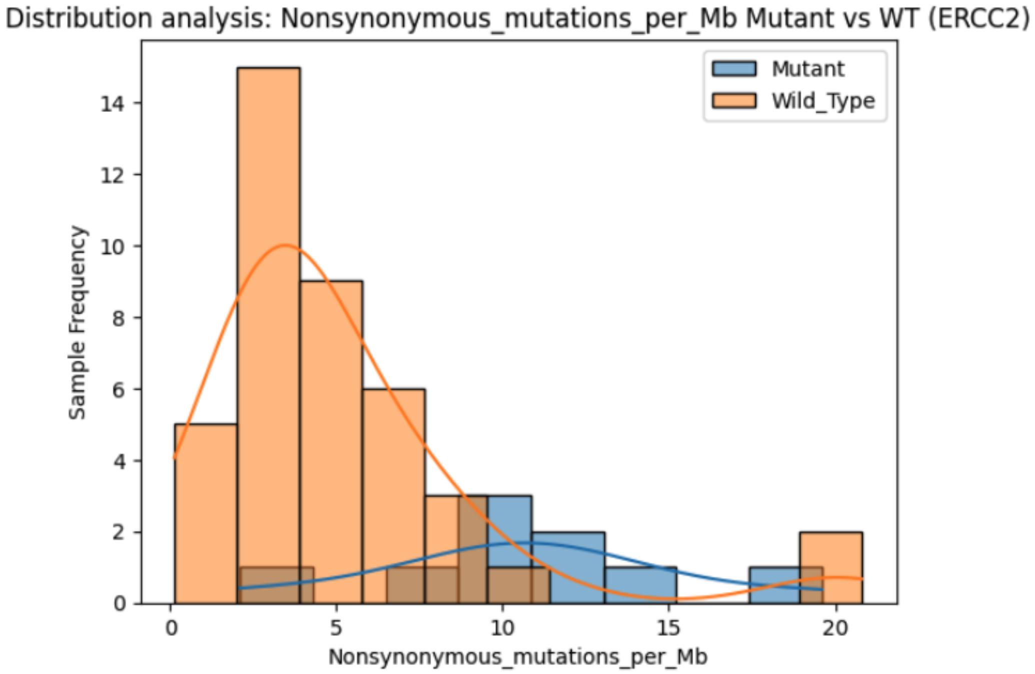


5. **How many samples are wild-type versus mutant with respect to the most significantly enriched gene from Question 4? Plot the number of nonsynonymous mutations per megabase in the mutant vs. wild-type samples. Is there a significant difference in the number of mutations between the two groups? Information on the number of nonsynonymous mutations per megabase for each patient can be found in the file data/sample-information.tsv.**

There are 9 responders with mutant ERCC2 gene and 41 non-responders. By Mann-Whitney U test we find a significant difference in the number of mutations between the two groups. (Although we would prefer if the shapes of curves were more similar, I tested after removing outliers as well in code)

The difference between the two groups based on non-synonymous mutations has been significant.





6. **Write any conclusions that you have made based on your analysis. How might this analysis be improved or expanded upon? Please include all requested figures in your report.**

- Use Bioconductor package like maftools when possible :D
- The VAF of both the groups, responders and non-responders is below 0.5. In diploid species ideally close to 0.5 expected for a medium stage tumor assuming purity and sequencing quality and region selectivity is good. Indiviual patient's VAF or smaller groups might give more specific insights and an understanding of the heterogenous mutations level. Based on the cancer type and location, such as colorectal cancer and with some more analysis, we may be able to comment that the tumor is in initial stages of development with Darwinian evolution due to considerable intratumoral heterogeneity, arising from multiple sub-clonal mutations at initial stages (Saito et al. 2018) https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6056524/

  We may be able to make a comment that a considerable portion of mutations are passenger mutations and hence are not being selectively favoured in general and resulting in low VAF in the sample population. The impact of treatment on the VAF of mutation at the gene level in the groups cannot be significantly associated. We can potentially form more hypothesis and conclusions.
- It can be concluded that the treatment resulted in enriching each of the 7 mutant genes in a subset of responder patients (statistically significant in sample population, responder vs non-responder), implying resistance generated by subclonal mutations, at the sample collection time, within those patients. The treatment's impact in reducing mutants at gene level was not significantly considerable for any of the patients. (Perhaps one should perform one-tailed t-test for this to strengthen argument)
- It would seem total mutation load post-treatment may have been one of the factors for classifying Response and Non-Response Patients for which the argument can be strengthened with a t-test/Mann-Whitney U test.
- In the gene analysis with most mutations and general presence per patient we found that these 5 genes SPEN, KMT2C, ZNF91, DST, TYRO3 are dominant in certain groups of patients and can be further evaluated for more information about any unique mutational signatures, potentially de-novo discoveries and personalized treatment recommendations.
- In statistical enrichment test post-treatment, we might be able to add some weightage to consider number of mutations level present in each gene in Response or Non-response patients. We could include at Variance Allele Frequency (VAF) to filter our data and also in our analysis at different levels such as at mutation per gene per patient per group. In some cases, it might even indicate potential germ-line mutations being present.
- ERCC2 was estimated to the most significantly enriched mutant gene being present in 9 Responders and 0 Non-Responders. It acts as a significant factor for the repair mechanism by provides instructions for making XPD protein, an essential part in TFIIH. Which is involved in transcription and repairing damaged DNA. Hence, a basis for multiple cancer types and a potential driver gene. (https://medlineplus.gov/genetics/gene/ercc2/) The number of mutant sample observed is 9 and remaining patients are dominantly considered to have wild-type copy of the gene, hence 41 wild-type samples.
- The difference in non-synonymous mutations in the two groups segregated by mutant and wild-type of ERCC2 gene (most significantly enriched gene) is statistically significant due to "$p < 0.05$". Further potential importance/significane of ERCC2 gene in non-synonymous mutation

load in patients can be done with one-tailed Mann-Whitney U test at the gene level (individual gene based mutations, link to other passenger and driver mutations in same and other genes, etc.) on different combinations of sample population based on requirements.

- There are still a number of analysis we could perform, including locus based mutations analysis to identify hotspots and driver mutations. We could perform functional and pathway analysis to understand the implications of mutations and their significance and other relevant mutations linked with them. Identifying gene-pair co-occurenece frequency. Perform analysis with mutation tables such as co-onco plots. Observe at entire protein change across sample population per gene and per patient/group and compare it with reference datasets. Using methods like Gene Set Enrichment Analysis to get analysis with a broader picture to analyze patterns and understand the functional significance of gene expression. Adding further clinical data. We could utilize mutational signatures and understand further about the cancer's potential evolution and detect its subtype. Utilize already referenced and filtered data such as from "TCGA" specific/similar/even contrasting to our case and identify new insights and hypothesis and accelerate our analysis and discovery.

  We could also perform further analysis based on germ-line (driver) mutations vs somatic (driver) mutations to further increase our predictive and causal analysis, by comparing germ (ideally) cells with tumor cell samples and adapt and improve our treatment results.

- Lastly, after having certain amount of data on sample population for different cancer types (such mutational signatures, mutation types, omics data and other important features), we need to shift from real-time situational analytics and basic predictive analytics to real-time advanced predictive analytics utilizing Artificial Intelligence (Machine Learning + Deep Learning + Natural Language Processing + Computer Vision + Reinforcement Learning + Geometric Deep Learning, and further advancements) enabling significant enhancement in knowledge and deeper insights, cheaper and automated base treatment solutions and analysis, treating relapses and clonal and sub-clonal driver mutations before they may even occur and several other benefits.

There's definitely much more that can be done.

Along with my analysis, deduction and technical skills, I hope you have found my potential to learn and think in-depth with the time I spent in this assignment, and it shows my commitment and ambition in improving our lives from cancer and other chronic diseases that humanity faces today.

Thanking you,
Hany Raza