האקטון 2022, משימה שנייה

מגישים:

בר נמש: 312162464. אביתר בן: 312245087. יובל עומר: 316514314.

## Pre-processing:

* The first challenge we encountered when working on the challenge were the medical terms And annotations which are prevalent in the data.

To face this challenge, we did extensive research on the internet.

Furthermore, at first glance it was noticeable that many features contain basic errors such as typos, and conflicting values, but most errors were recoverable and we attempted to turn them into useful data.

Similarly, we spent a lot of time thinking about each feature - in which way to process its values, Is it categorical? Is it ordered? How to deal with features whose values are dates? Are there new features that we can create from existing features? Which features seem to be of great importance and which are dispensable? What is the correlation[[1]](#footnote-2) between the different features?

Some examples:   
One of the features we encountered is the KI67 protein which is a protein that effects cell growth and proliferation, and many studies show it has a direct connection to cancer growth.  
Which led us to decided to categorize the values into a group of five categorical values- from low value to high value, this process required us to extract the correct value from a given input in spite of the many formats the input can be given.

Regarding dates features we conclude that the right way to deal with them is to calculate the difference between them – days amount between diagnosis and surgery for example.

* The data is made of the following features:
* Dates, patient's age, tumor size,
* We tried various methods that emits different results:

Different hyper parameters in different classifiers , such as Random Forest, Decision Tree,

KNN. In addition to that each classifier was observed in different multi-class classifier transformer, such as Power Set, Chain Model, and Binary model.

* We carefully learned our data, and take extra care to learn it without over-fit the data.

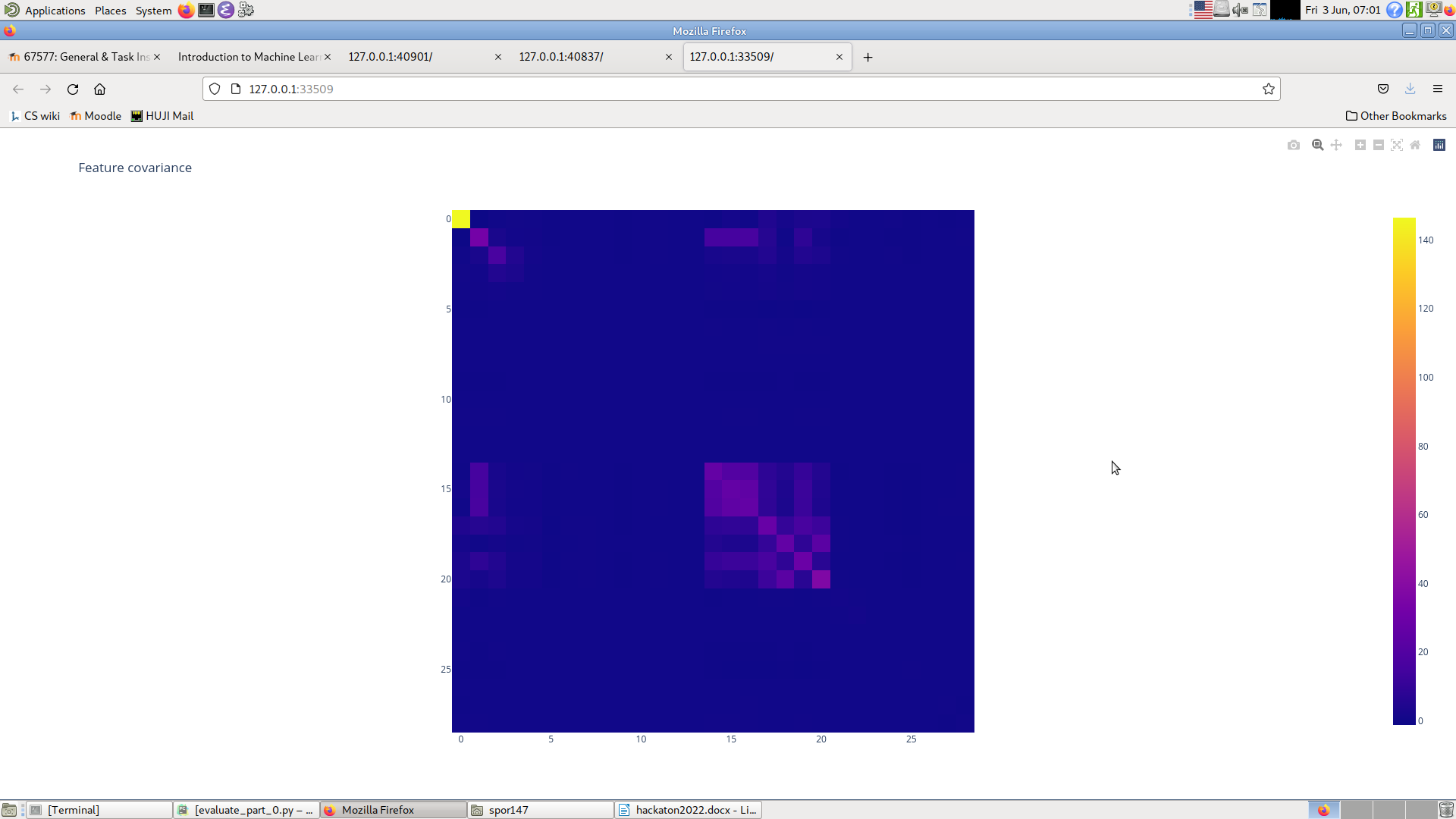
In order to do that we split our data to train and test parts, and we learned our data without peak at the test part until the very end.

In order to understand about the variance and bias that exists in our model we re-sampled our train data using the cross-validation method.

Unfortunately, as we mention in the next part, we failed to properly process the data, and missed duplicated values.

* We ended up using a Classifier Chain transforming a Decision Tree classifier, with little regularization and cv, as it was the one that emits the better results.

Figure 1: Correlation between features



## Part 3:

Interesting things we found in the Hackathon:

We created two figures which we think can show some intresting aspects of our data-set

The first[[2]](#footnote-3) is a Figure showing our data after a projection on distanes from 3 Kmeans clusters.

In this figure we can clearly see that the clustering did a pretty good job of differentiating age groups, as seen by the continuous color change between sides of the graph.

In the second[[3]](#footnote-4) graph we can see that the principal components of our data aligns with the stages of cancer, as seen by the layers in the graph.

The third is a graph showing Age over time between the diagnosis and the first surgery which is in the shape of a Gaussian distribution.

All graphs can be recreated in our code using a simple terminal command.

Figure 2: Kmeans clustering with age as color

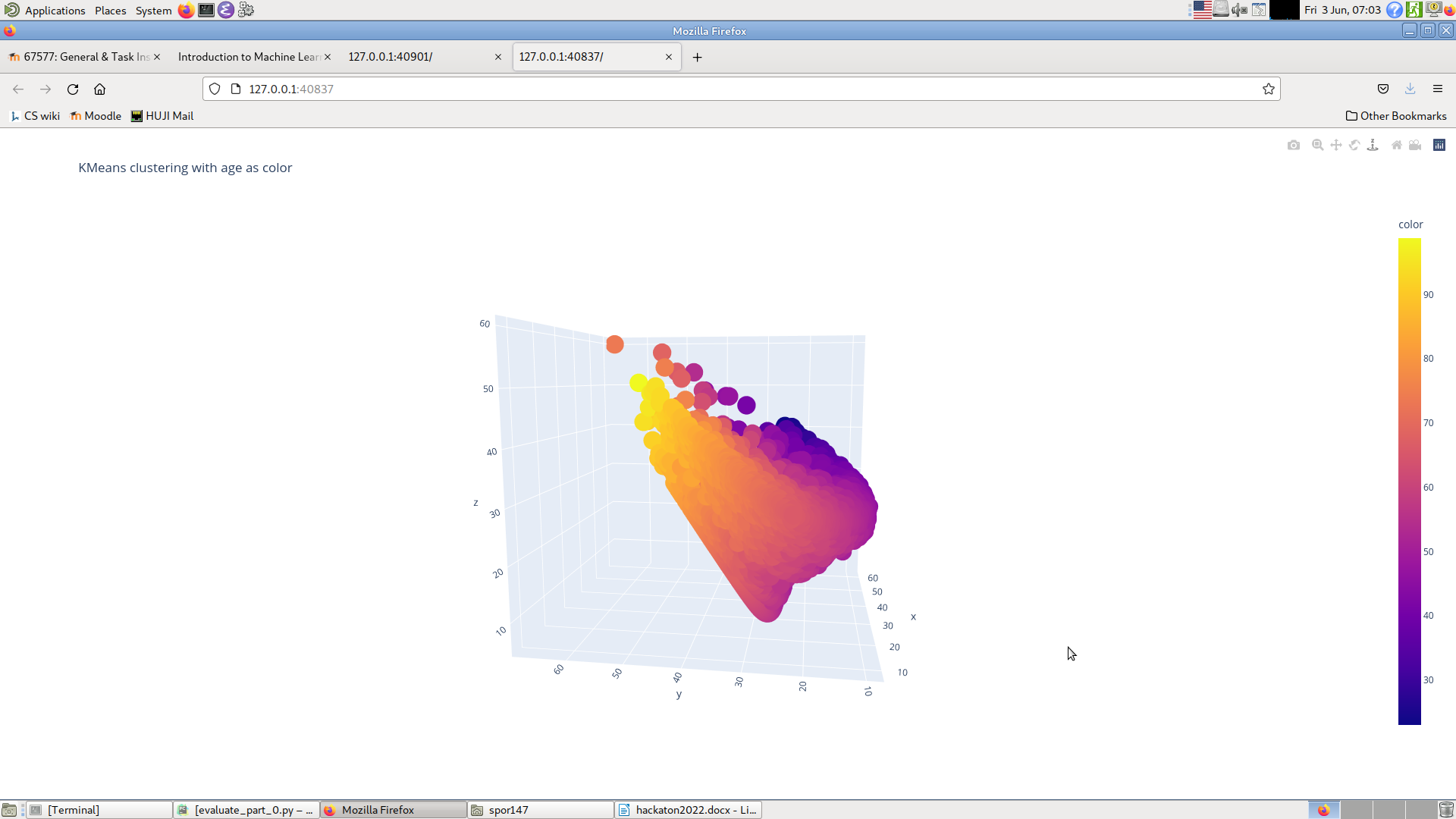


Figure 3: PCA components analysis over cancer stage

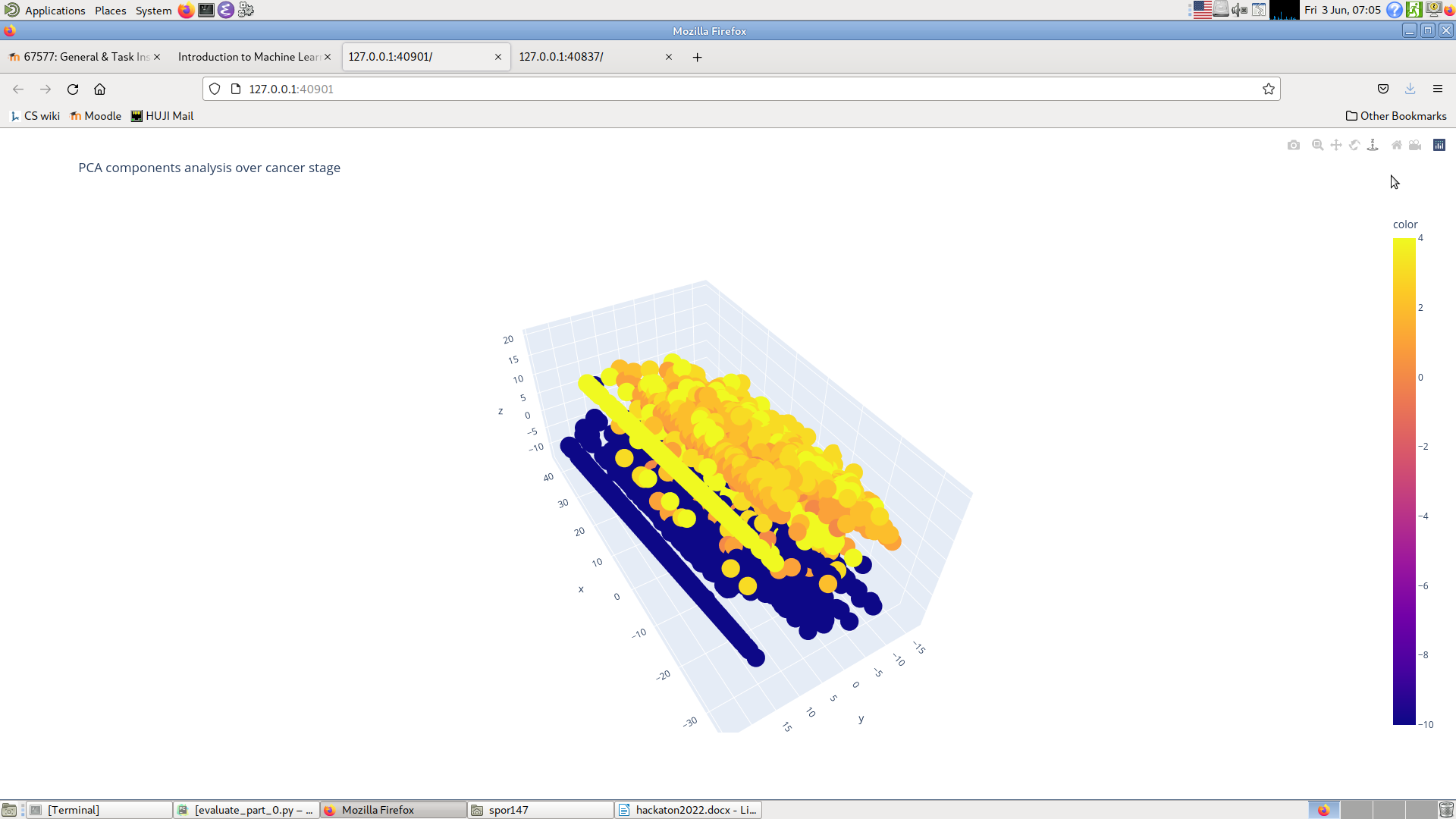
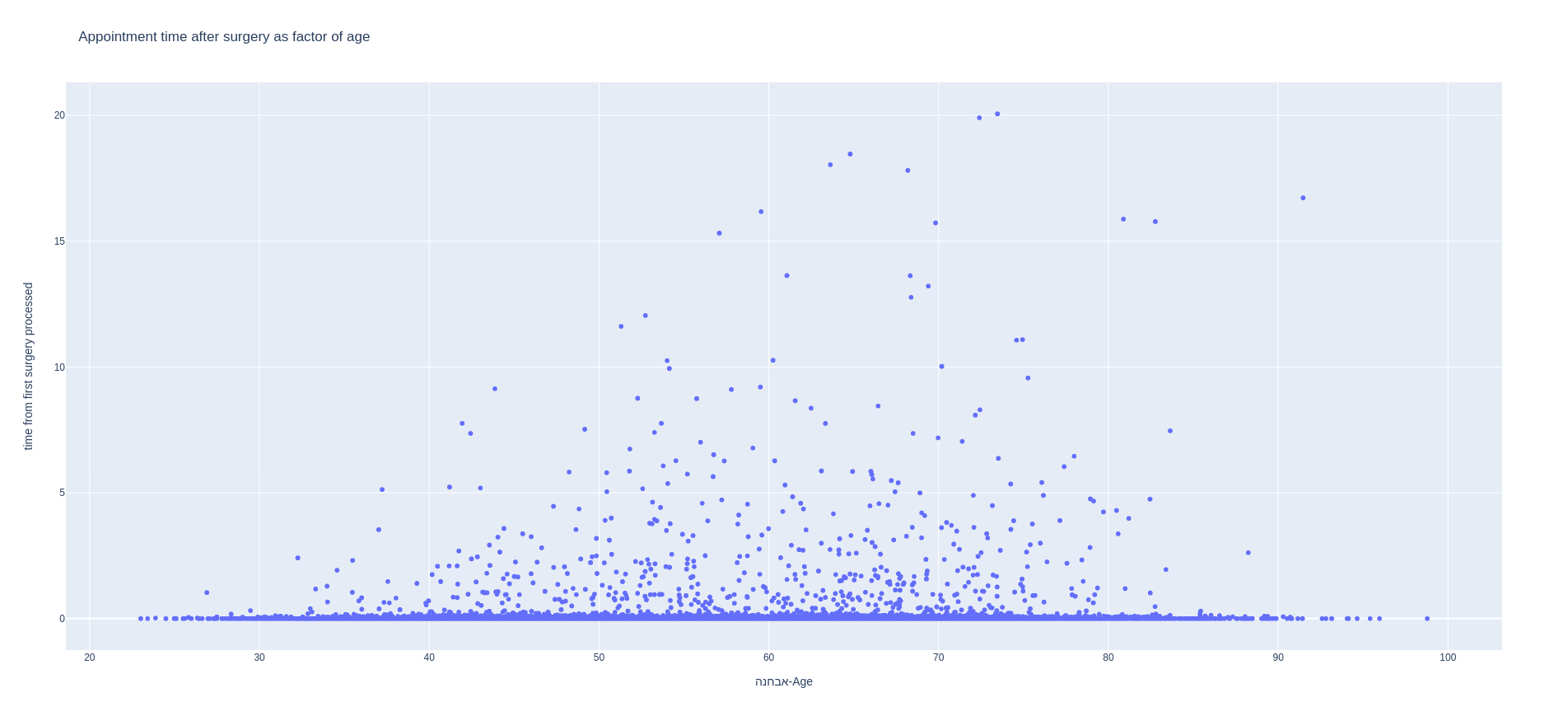


Figure 3: Age over time between diagnosis feature and first surgery in day.



Conclusions From this task:

This task was both intriguing and challenging and we have learned many things from trying to solve it.  
In the end we didn’t manage to train a model that learns the underlying aspects of this task but we did leave with a few insights on how we should have approached this task better.

First and foremost our examination of the data wasn’t thorough enough, we didn’t notice until a relatively late stage of the task that there are many duplicate samples ,that differed mainly by the doctor entering the data and the form entered, which should be dropped in pre-processing in order to avoid over-fit.

Our pre-processing was lackluster, as evident by the poor performance of our model, we should have spent more time on building our data foundation since a strong model needs a stronger data-set.

Furthermore due to lack of time we didn’t do EDA properly which prevented us from deriving more features from our original data.

In the end we learned a lot from this experience and even had fun along the way.

Thank you for giving us the opportunity to partake in this Hackathon!

1. See figure 1 [↑](#footnote-ref-2)
2. Figure 2 [↑](#footnote-ref-3)
3. Figure 3 [↑](#footnote-ref-4)