Narek Koucherian

004780315

BE175 Final Project Report

“Extremely High Mutation Rate of HIV-1 in Vivo”

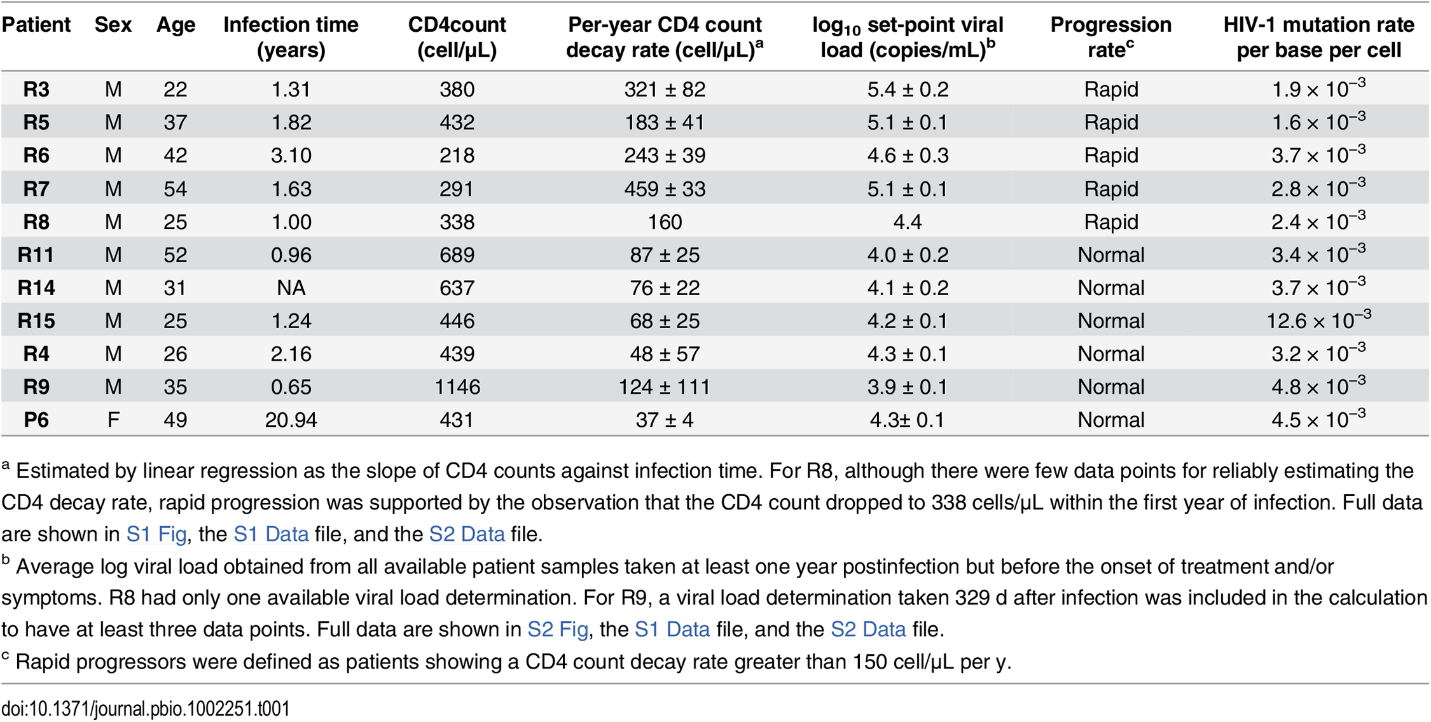
José M. Cuevas, Ron Geller, Raquel Garijo, José López-Aldeguer, Rafael Sanjuán

**Introduction/Motivation**

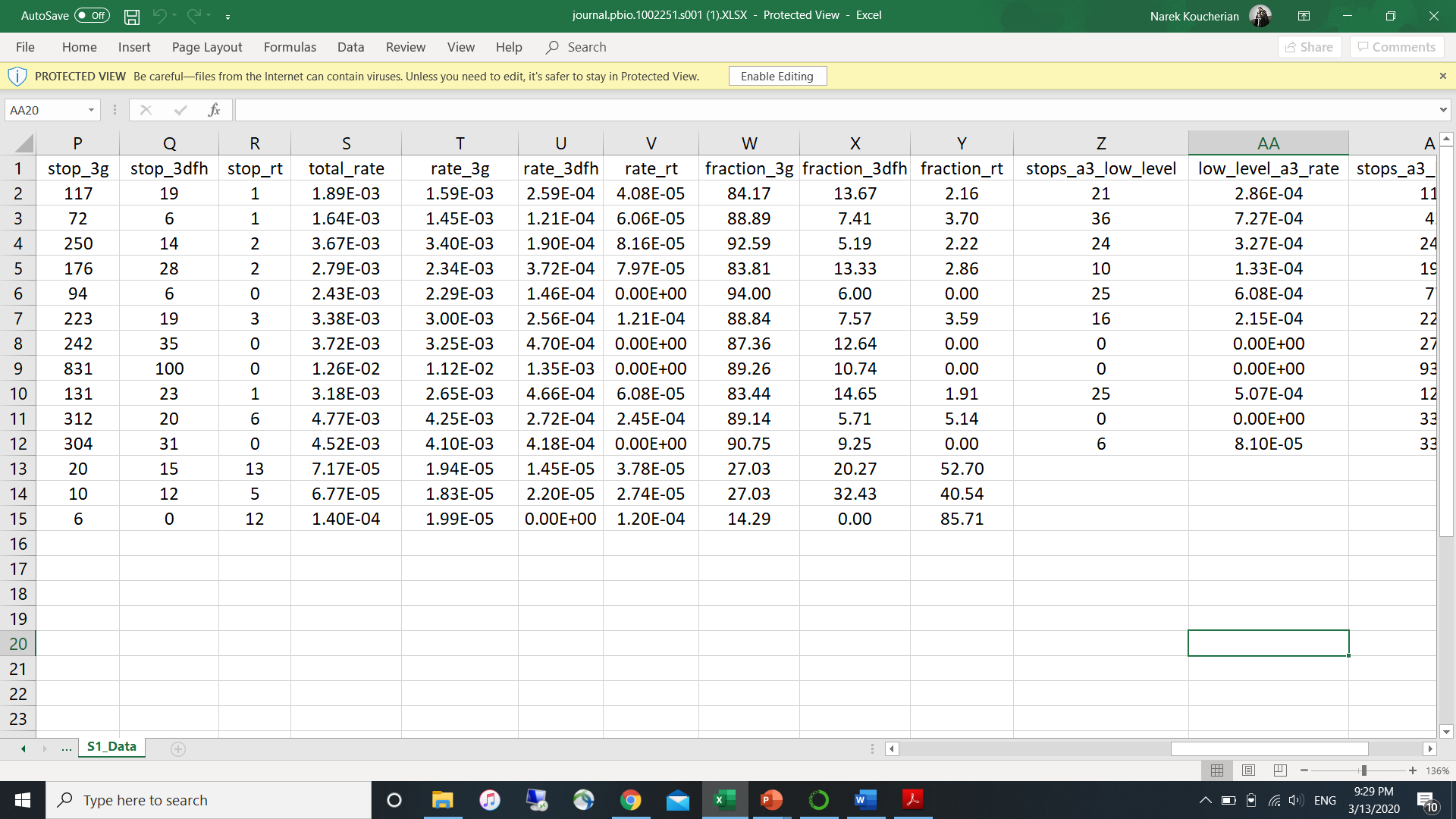
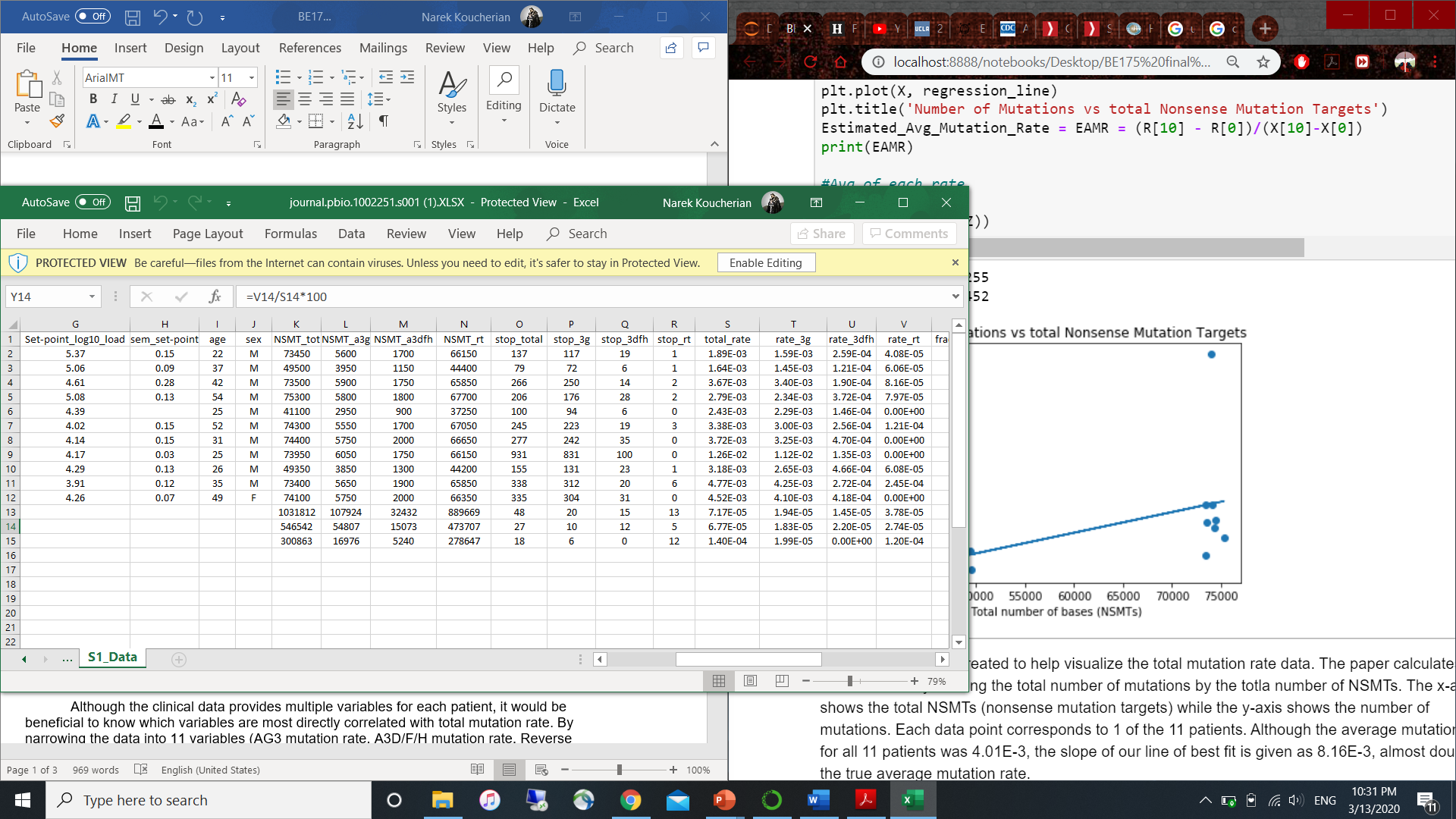
Human Immunodeficiency Virus (HIV) is a virus that weakens the immune system by targeting and destroying T-Helper cells, also known as CD4 cells. The virus cannot replicate on its own, so it transfers it viral RNA into CD4 cells and converts them into DNA. This creates new copies of the virus allowing it to spread and infect more cells. When the hosts CD4 cell count diminishes past 200 cells per microliter, the host is classified as having Acquired Immunodeficiency Syndrome (AIDS).1,2

The HIV epidemic began in 1981, infecting over 75 million people within the past 40 years. A 2019 study from UNAIDS place the worlds infected population at 37.9 million, with an estimated 20% of the infected population not being aware that they have HIV.4 Due to the virus’s high mutation rate, no vaccine has been developed to successfully combat HIV. The most advance method of treatment for HIV is Antiretroviral Therapy (ART). This method involves patients taking antiretroviral drugs to disrupt the various point of HIV’s replication process. ART allows people living with HIV to decrease their viral load, strengthen their immune systems, and cope with the side effects of the virus.5

Cuevas et al approach the issue of HIV-1 by studying the mutation rates associated with the virus in vivo. Their study examined 11 patients with HIV and classified their total HIV-1 mutation rate per base cell by studying sequences within Peripheral Blood Mononuclear Cells (PBMCs). Mutations were identified by counting the number of premature stop codons in each sequence and dividing them by the total Nonsense Mutation Targets (NSMTs). The study also provides additional clinical data for each patient including but not limited to age, CD4 count, CD4 decay rate, viral load, rates for 3 enzyme specific mutations, and fractional compositions of each enzyme mutation rate within the total mutation rates.3



**Figure 1** – Clinical data for the 11 patients. Patient codes are listed in the left column, followed by sex, age, duration of infection in years, CD4 count, per-year CD4 decay rate, log10 set-point viral load, progression rate, and total mutation rate per base cell. Patient R14’s infection time data was not available and was artificially set as 1 year for the purpose of this project.3

**Figure 2A**- Additional data pulled this paper’s data. Includes total mutation rate (same as last column of **Figure 1**), rate of A3G enzyme mutation, rate of A3D/F/H enzyme mutation, rate of Reverse Transcriptase enzyme mutation, fraction of A3G mutation rate, fraction of A3D/F/H mutation rate, and fraction of Reverse Transcriptase mutation rate. **2B** – The data on the right includestotal number of Nonsense Mutation Targets. NSMTs are categorized as being attributed into the 3 enzyme subtypes. The rightmost column shows the total number of premature stop codons used to quantify the number of mutations per sequence.3

Two unnamed students from Professor Meyer’s BE188 course have reimplemented the data from this paper before. While their work focused on developing a new algorithm for calculating the mutation rate of each patient, this new project focuses on implementing Partial Least Squares Regression (PLSR) to analyze the data.

**Problem Definition**

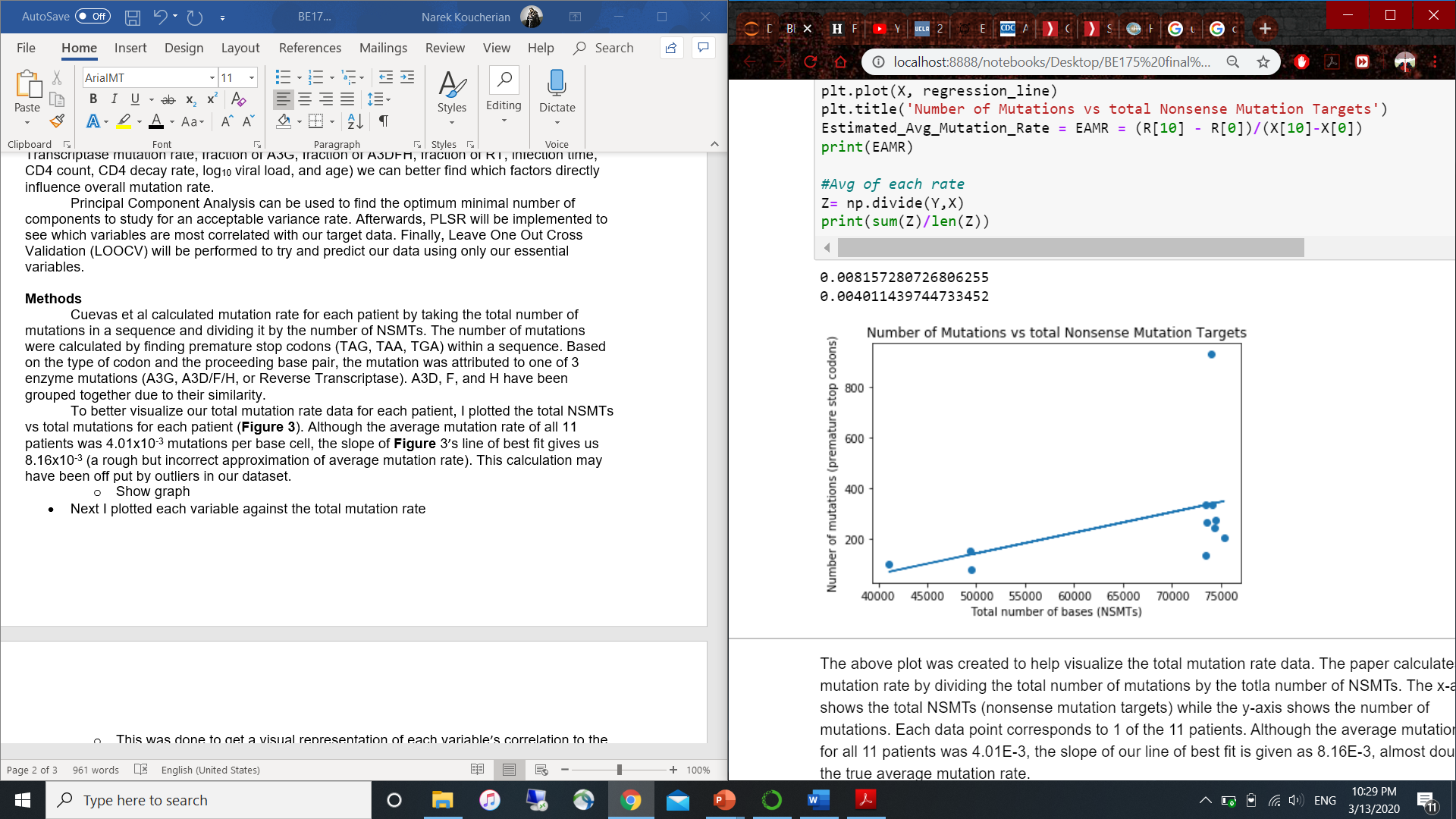
Although the clinical data provides multiple variables for each patient, it would be beneficial to know which variables are most directly correlated with total mutation rate. By narrowing the data into 11 variables (AG3 mutation rate, A3D/F/H mutation rate, Reverse Transcriptase mutation rate, fraction of A3G, fraction of A3DFH, fraction of RT, infection time, CD4 count, CD4 decay rate, log10 viral load, and age) we can better find which factors directly influence overall mutation rate.

Principal Component Analysis (PCA) can be used to find the optimum minimal number of components to study for an acceptable variance rate. Afterwards, PLSR will be implemented to see which variables are most correlated with our target data. Finally, Leave One Out Cross Validation (LOOCV) will be performed to try and predict our data using only our essential variables.

**Methods/Results**

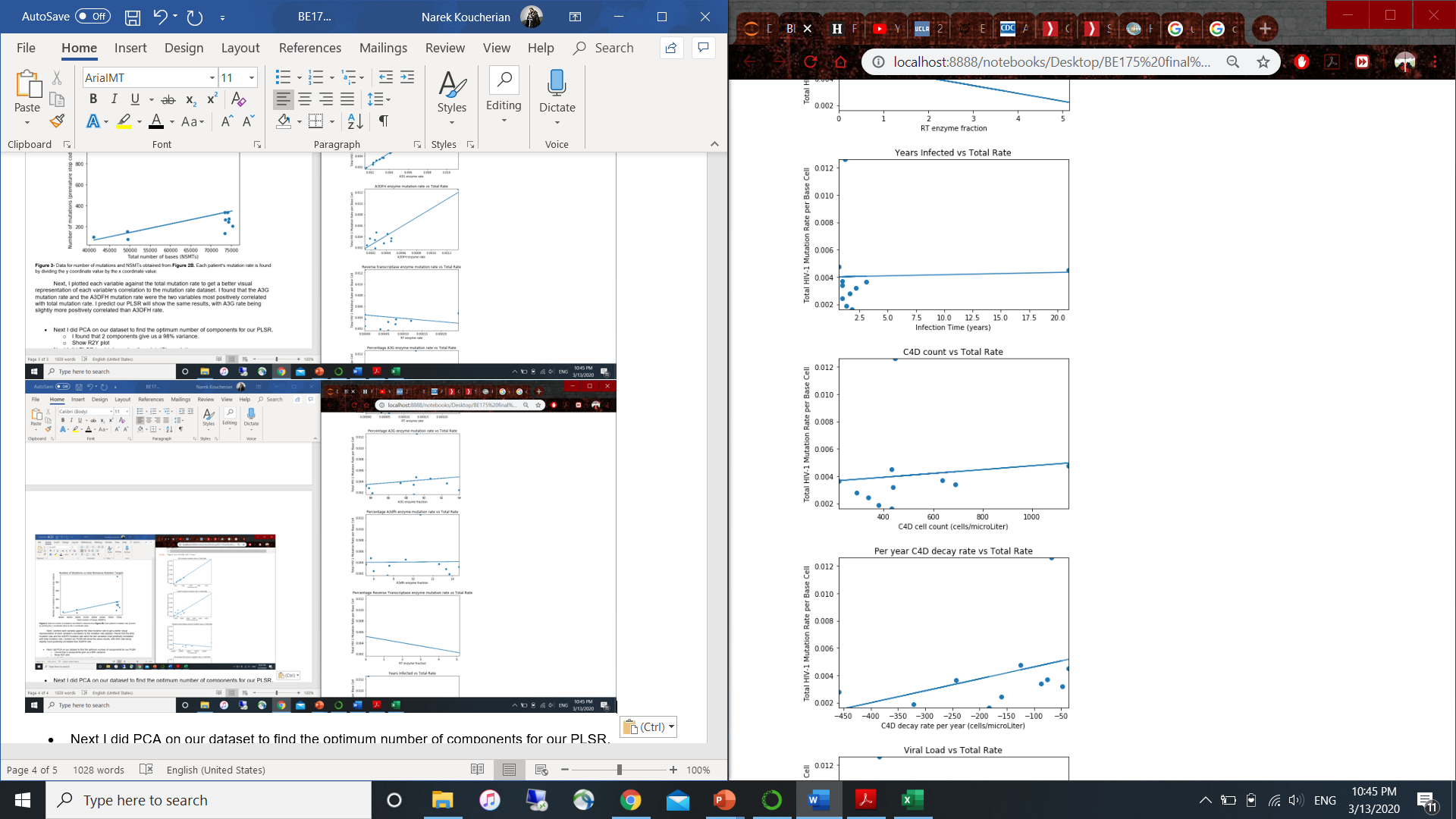
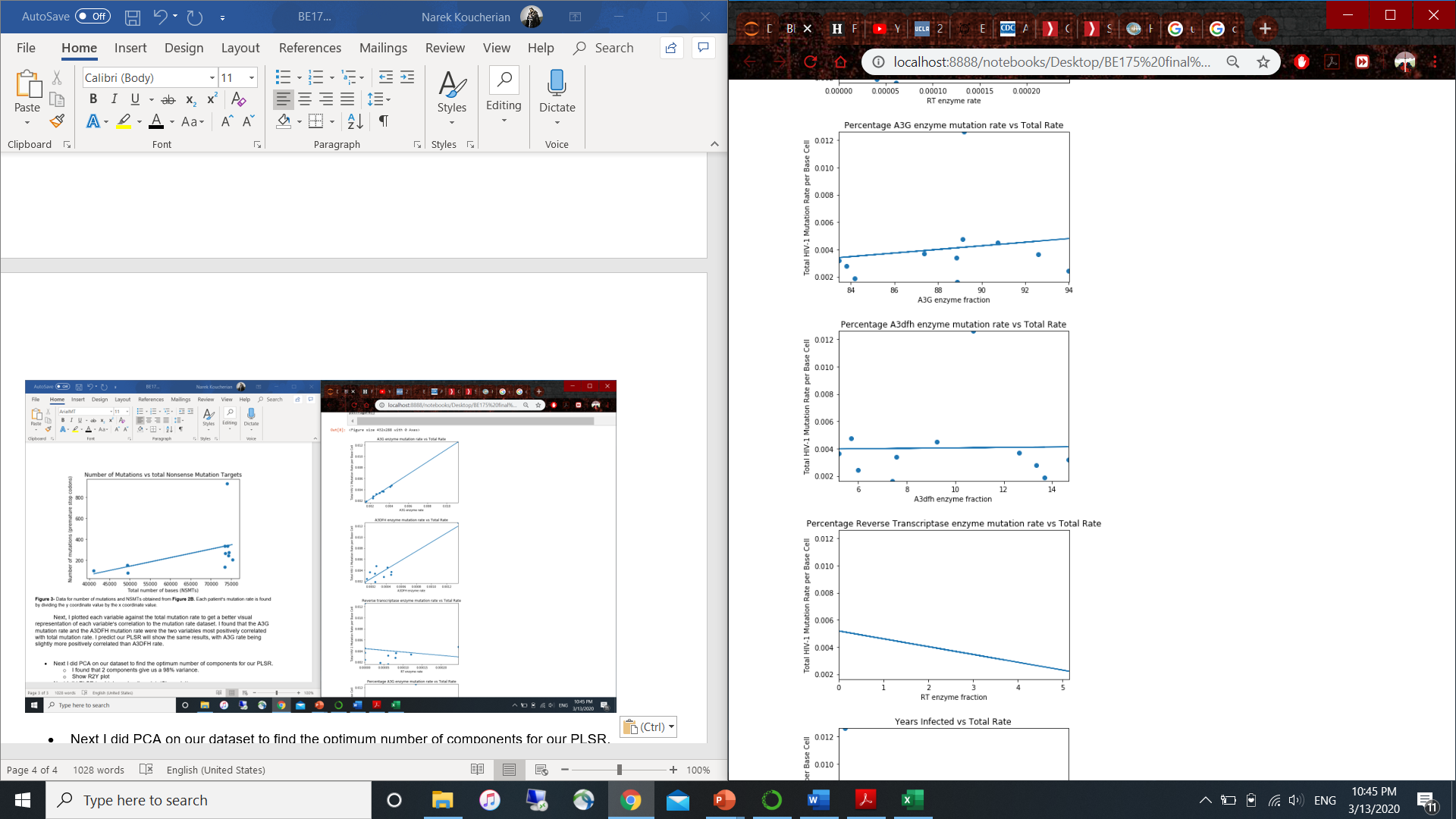
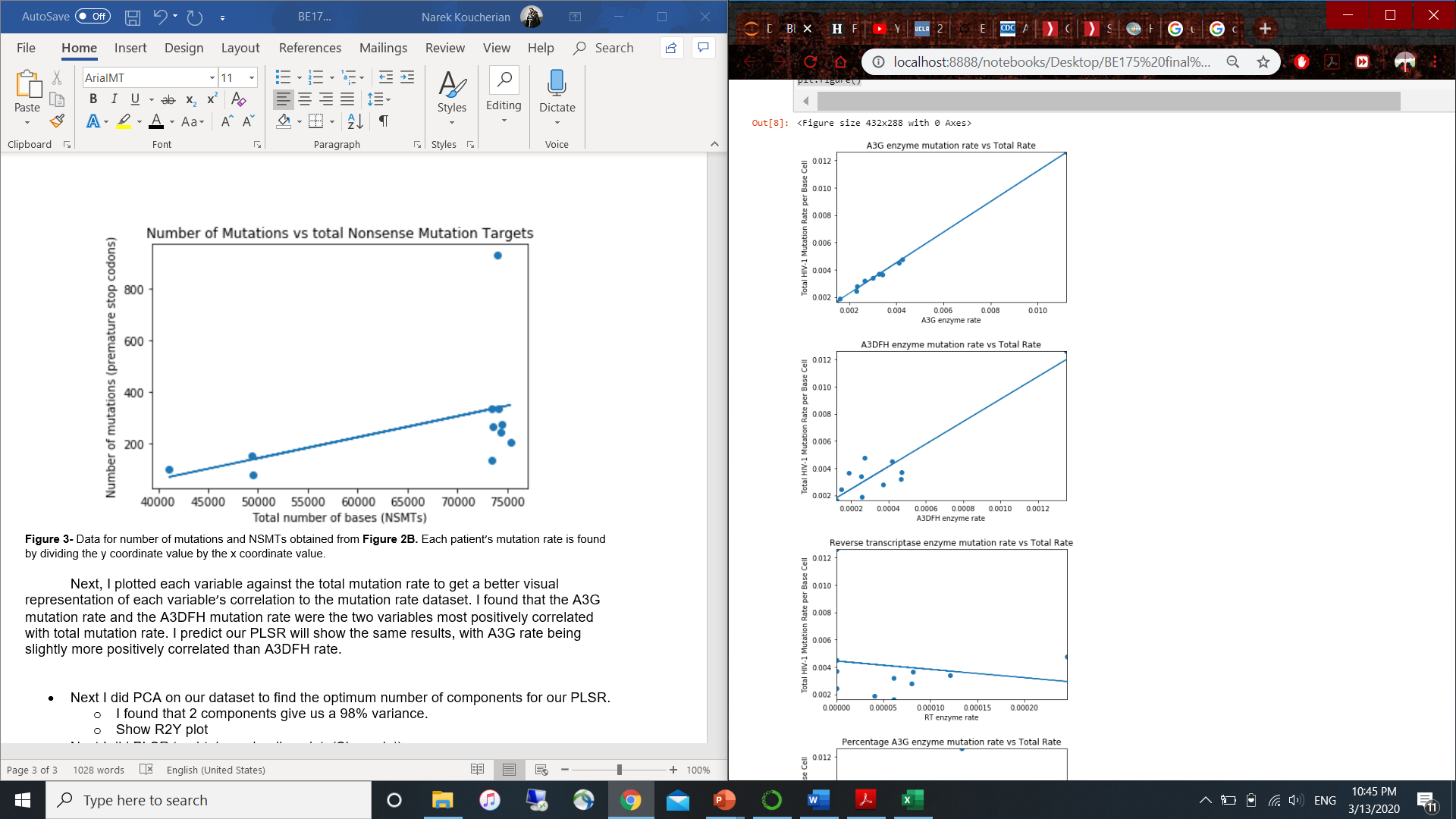
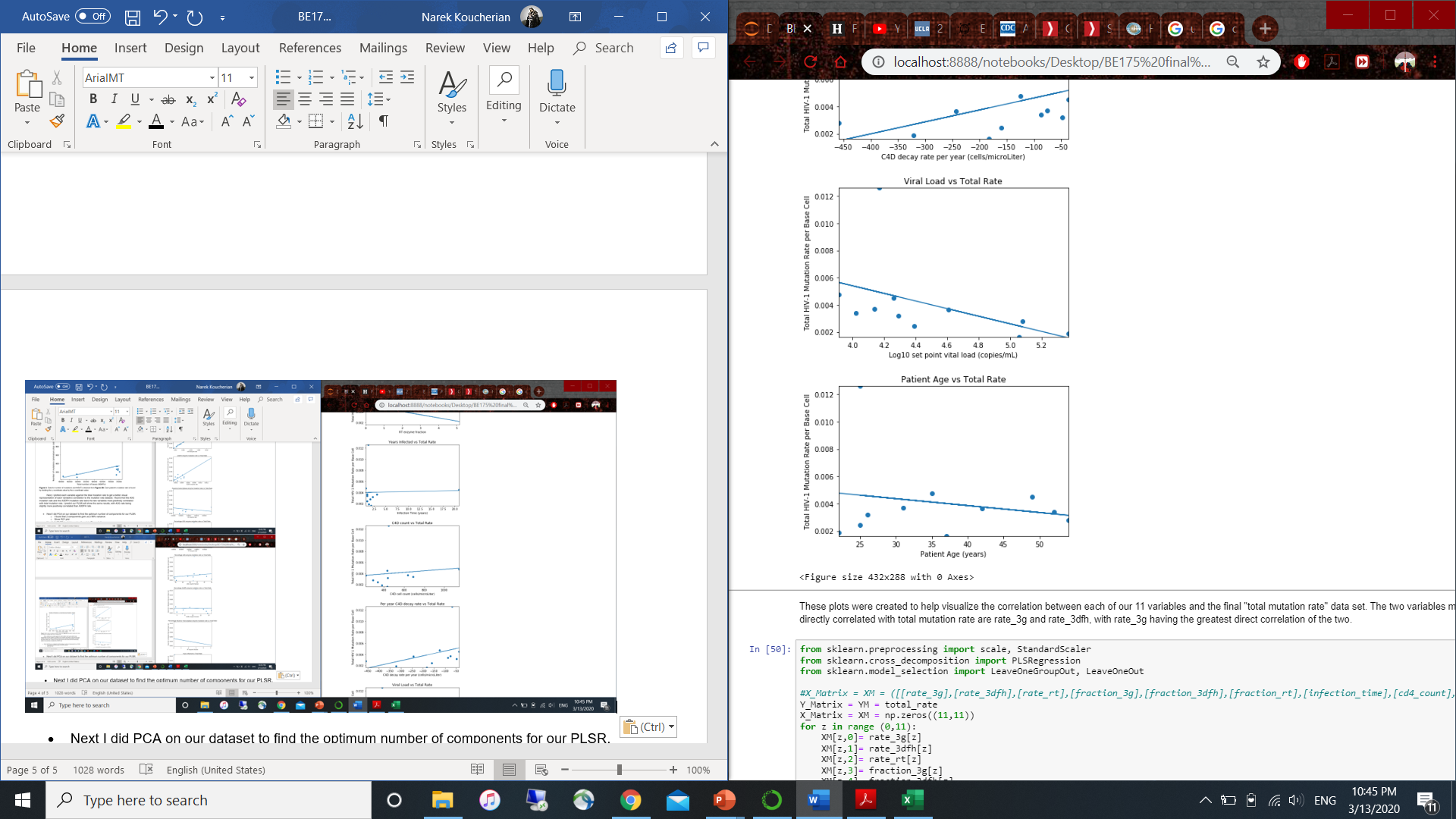
Cuevas et al calculated mutation rate for each patient by taking the total number of mutations in a sequence and dividing it by the number of NSMTs.The number of mutations were calculated by finding premature stop codons (TAG, TAA, TGA) within a sequence. Based on the type of codon and the proceeding base pair, the mutation was attributed to one of 3 enzyme mutations (A3G, A3D/F/H, or Reverse Transcriptase). A3D, F, and H have been grouped together due to their similarity.3

To better visualize our total mutation rate data for each patient, I plotted the total NSMTs vs total mutations for each patient (**Figure 3**). Although the average mutation rate of all 11 patients was 4.01x10-3mutations per base cell, the slope of **Figure** 3’s line of best fit gives us a value of 8.16x10-3 (a rough but incorrect approximation of average mutation rate). This calculation may have been off put by outliers in our dataset.



**Figure 3-** Data for number of mutations and NSMTs obtained from **Figure 2B.** Each patient’s mutation rate is found by dividing the y coordinate value by the x coordinate value.

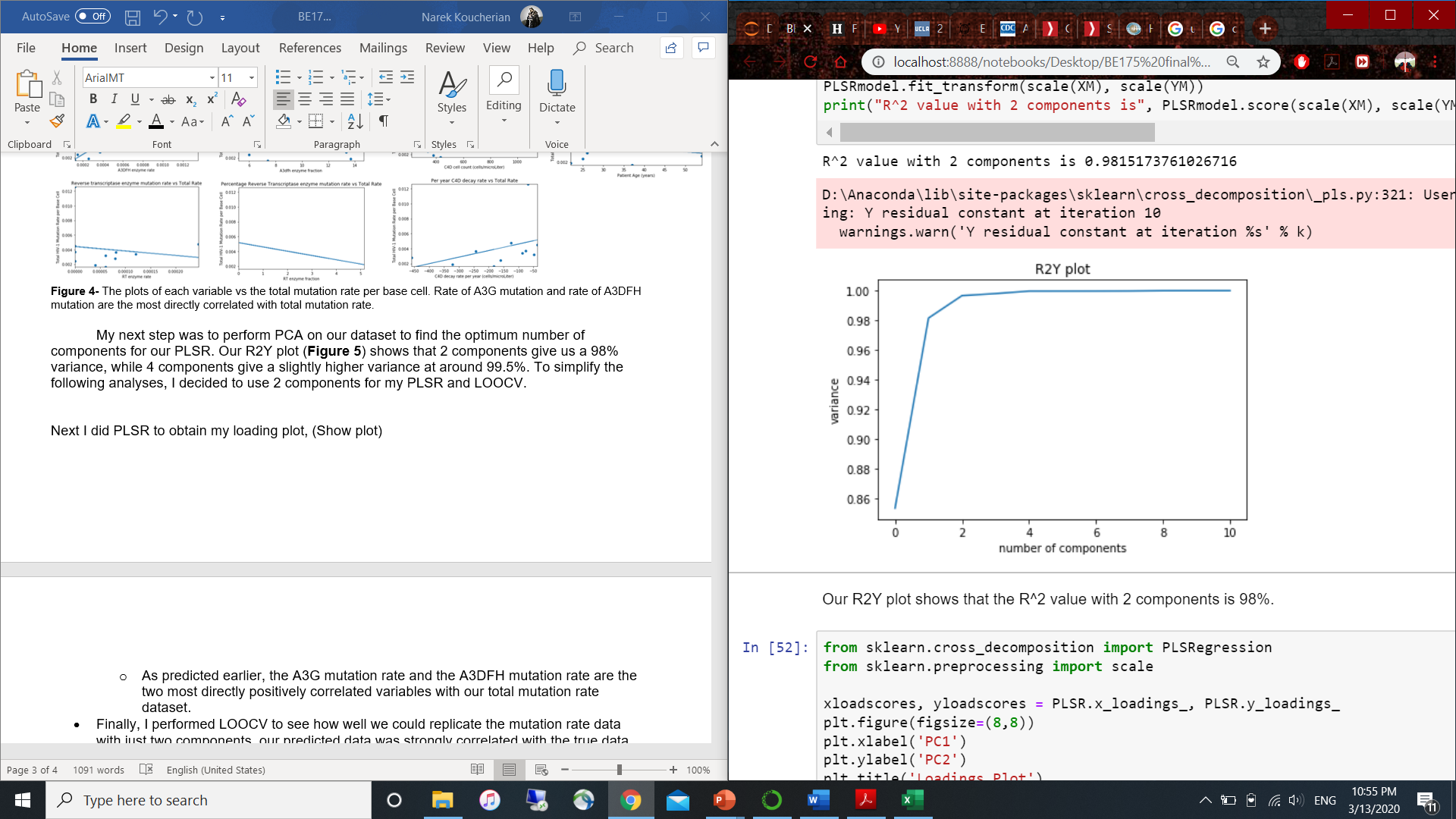
Next, I plotted each variable against the total mutation rate to get a better visual representation of each variable’s correlation to the mutation rate dataset.I found that the A3G mutation rate and the A3DFH mutation rate were the two variables most positively correlated with total mutation rate. I predict our PLSR will show the same results, with A3G rate being slightly more positively correlated than A3DFH rate.



**Figure 4-** The plots of each variable vs the total mutation rate per base cell. Rate of A3G mutation and rate of A3DFH mutation are the most directly correlated with total mutation rate.

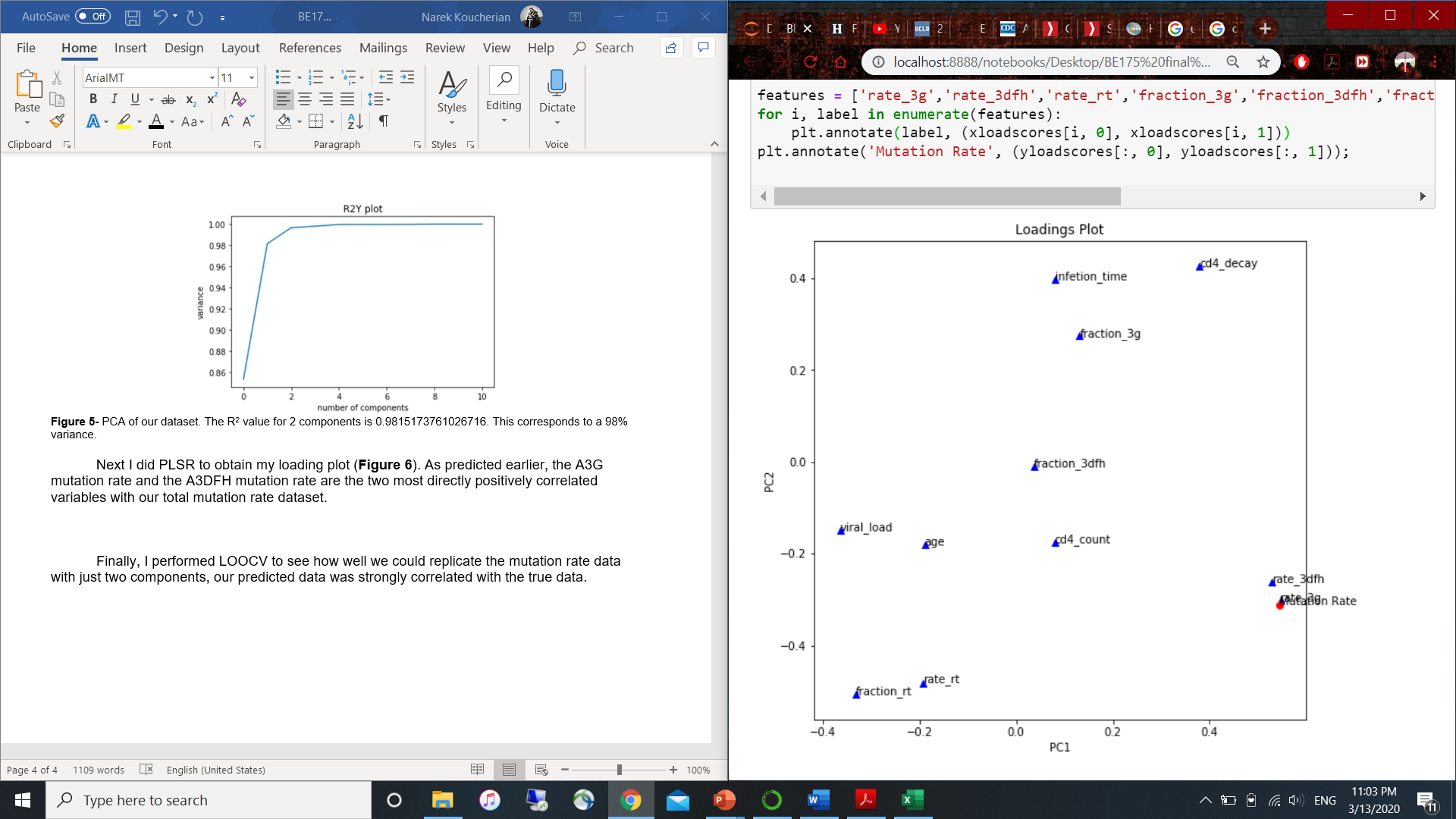
My next step was to perform PCA on our dataset to find the optimum number of components for our PLSR.Our R2Y plot (**Figure 5**)shows that 2 components give us a 98% variance, while 4 components give a slightly higher variance at around 99.5%. To simplify the following analyses, I decided to use 2 components for my PLSR and LOOCV.

­



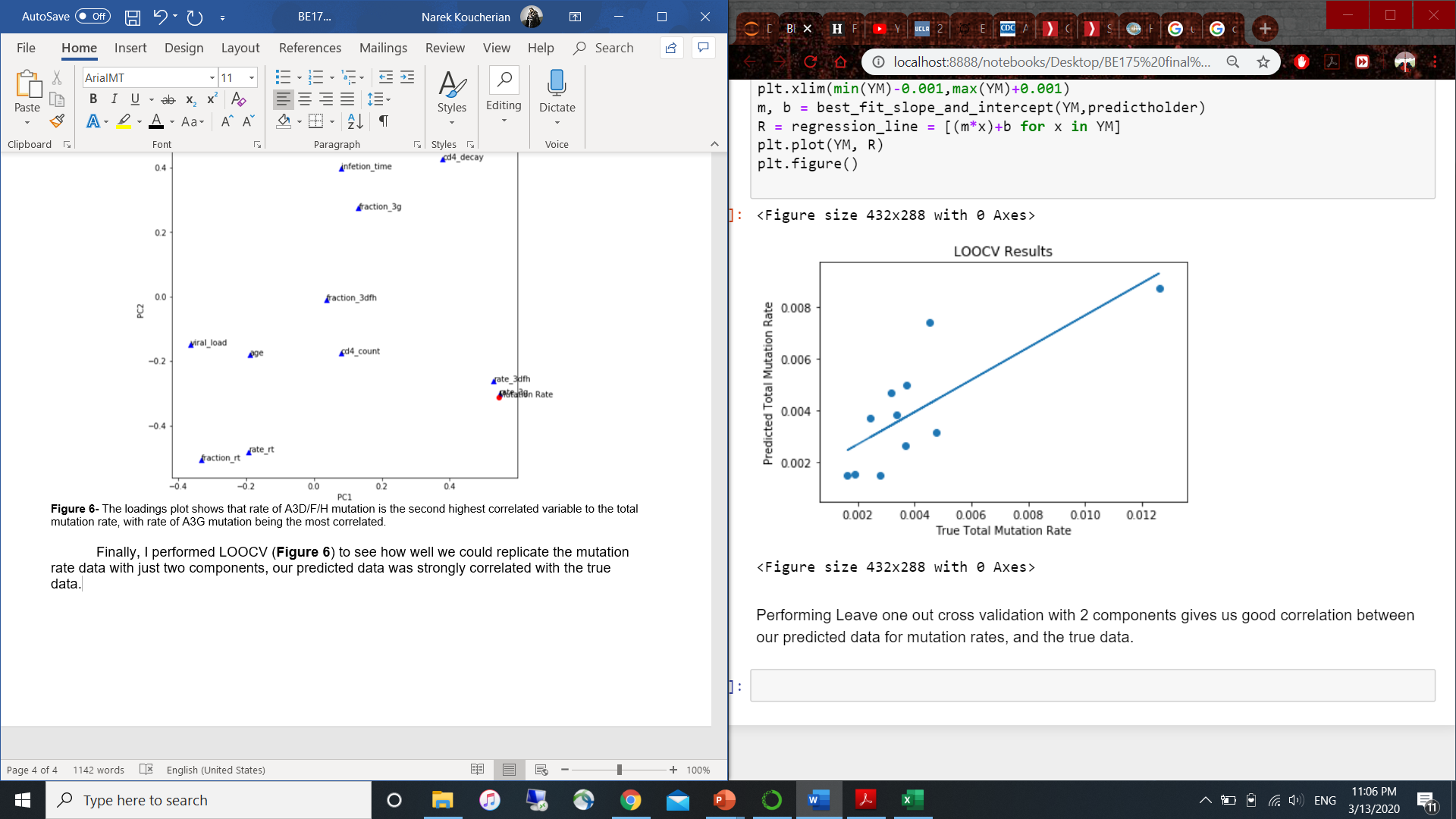
**Figure 5-** PCA of our dataset. The R2 value for 2 components is 0.9815173761026716. This corresponds to a 98% variance.

Next I did PLSR to obtain my loading plot (**Figure 6**).As predicted earlier, the A3G mutation rate and the A3DFH mutation rate are the two most directly positively correlated variables with our total mutation rate dataset.



**Figure 6-** The loadings plot shows that rate of A3D/F/H mutation is the second highest correlated variable to the total mutation rate, with rate of A3G mutation being the most correlated.

Finally, I performed LOOCV (**Figure 7**) to see how well we could replicate the mutation rate data with just two components, our predicted data was strongly correlated with the true data.



**Figure 7-** The results of Leave One Out Cross Validation performed on our dataset.

Overall, we found that the rate of A3G enzyme mutation and the rate of A3D/F/H enzyme mutation are highly correlated with total mutation rate. By focusing efforts on studying these two factors, scientists can accelerate their research in HIV-1 and increase efficiency in their efforts to develop a vaccine.

Works Cited

1. “About HIV/AIDS.” *Centers for Disease Control and Prevention*, Centers for Disease Control and Prevention, 2 Dec. 2019, www.cdc.gov/hiv/basics/whatishiv.html.
2. Admin. “How HIV Infects a Cell.” *International Partnership For Microbicides*, 10 Feb. 2016, www.ipmglobal.org/how-hiv-infects-cell#:~:text=When%20HIV%20infects%20a%20cell,on%20to%20infect%20other%20cells.
3. Cuevas, José M., et al. “Extremely High Mutation Rate of HIV-1 In Vivo.” *PLOS Biology*, Public Library of Science, journals.plos.org/plosbiology/article?id=10.1371%2Fjournal.pbio.1002251#pbio.1002251.s011.
4. “Global HIV and AIDS Statistics.” *Avert*, 18 Feb. 2020, www.avert.org/global-hiv-and-aids-statistics.
5. “Starting Antiretroviral Treatment for HIV.” *Avert*, 4 Mar. 2020, www.avert.org/living-with-hiv/starting-treatment.