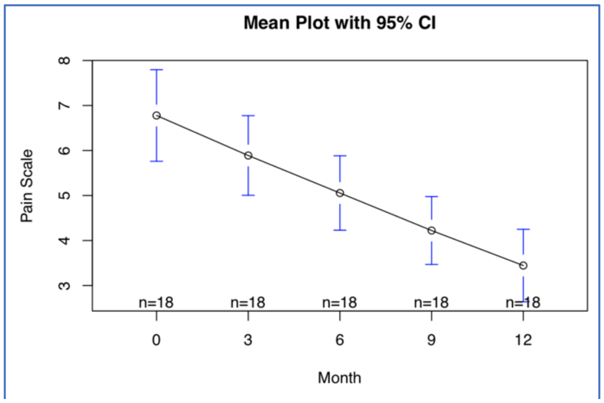
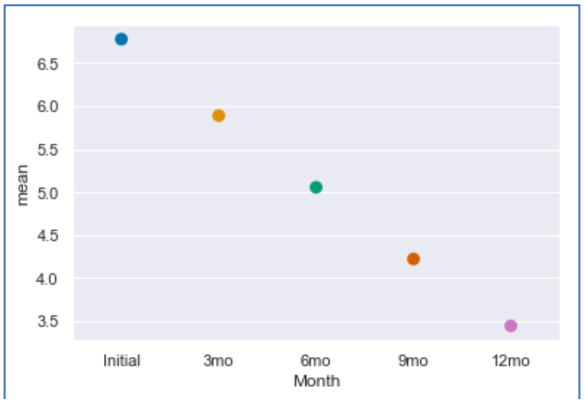
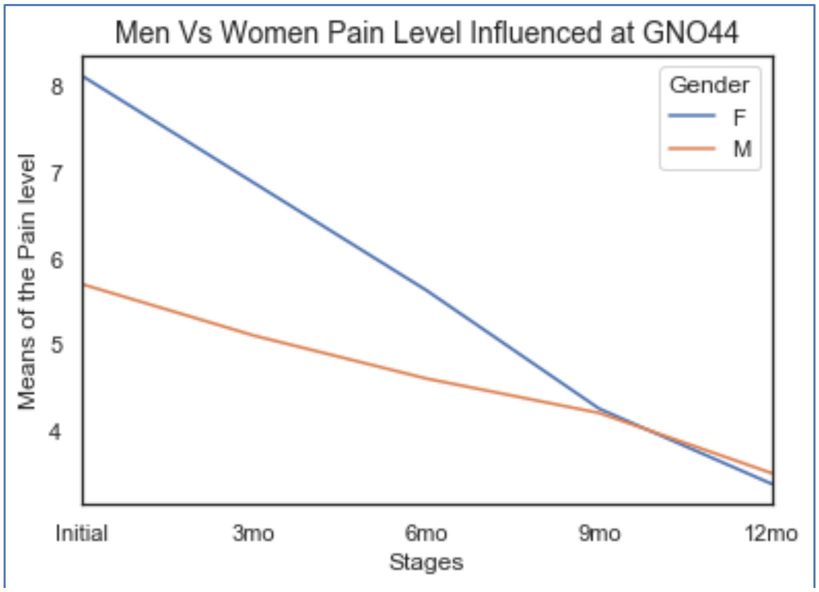
1, Our goal is to analyze the influence of the GNO44 on the reducing the perceived pain due to the arthritis. We are given the predictors: scare of pains at during period stages: beginning (0 month), 3 months, 6 months, 9 months, 12 months.

One-way ANOVA (analysis of variance) on different stages pain level is performed to figure out if there is a significant change in the pain level at different stages. The ‘F-test’ demonstrated a significant changes in the means of pain level at different stages. In othe words, the means of pain at different stages are spread out further from the overall mean, their variance is higher.





Stages	Mean	STD
Initial (0month)	6.777778	2.045240
3 Month	5.888889	1.778595
6 Month	5.055556	1.661757
9 Month	4.222222	1.516791
12 Month	3.444444	1.616904



Gender	0 Month	3 Month	6 Month	9 Month	12 month
Female	8.125	6.875	5.625	4.25	3.375
Male	5.700	5.100	4.600	4.2	3.5

I groupby the mean based on the gender cause I was curious about whether gender will experience different level of the pain level. In fact we see between beginning and 9 month, Female has bigger pain level drop comparing to Men. If we are given more data about different potential predictors: Age, job position, history of the illness and etc. we can discover more impact of the GNO44.

2,

We fit all of the predictors inside the model, and then we perform wald test to test the difference of the coefficients between: 'ddC' and 'ddI'.

I grouped the data into two groups: death and alive. For death group, the length of the time if the length of the life. In both death and alive groups, I analyzed the effect of the DDC and DDI based on the comparison of the length of the time. I used one-way between group ANOVA.

P-value of Drug-DDI	
Alive Group	Death Group
0.62	0.782

In both groups, I didn't see significant difference between DDC and DDI. The DDC has been omitted and the coefficient of the DDI (i.e. 0.2478) denotes the difference between the coefficient of the DDI and corresponding (i.e. DDC) level. So the difference between the coefficient of DDI and DDC woud be 0.24782. Since the P value for the DDI coefficient is not statistically significant. So I wouldn't say that there is a significant effect between drug DDI and DDC.

Predictor	Estimate Coeff	Std Error	Pr
Drug-DDI	0.24782	0.28179	0.379
Intercept	6.8722	0.73127	< 2e-16
CD4	-0.17702	0.03573	7.27e-07
Time	-0.49529	0.05183	< 2e-16

To check whether variable 'drug' improves the model fit, we fit the model with 'drug' variable and on second model without variable 'drug'. After conducting the likelihood ratio test, the 'drug' variable is not significant due to the P-value >0.05.

Model 1: Death~Drug+CD4+Time
Model 2: Death~CF4+Time
Hypothesis testing: Assume the model 2 is a much better and accurate model.

Predictor	Resid DF	DF	DF deviance	Pr(>Chi)
Model 1	463	323.11	N/A	N/A
Model 2	464	323.89	-0.77477	0.3787

3, This problem is a logistic regression problem. Since response is a binary variable: whether the patient has survived five years past diagnosis. Logistic regression model has been applied to study whether the predictor is risk enough for death prior to five years of a Hodgkin's diagnosis.

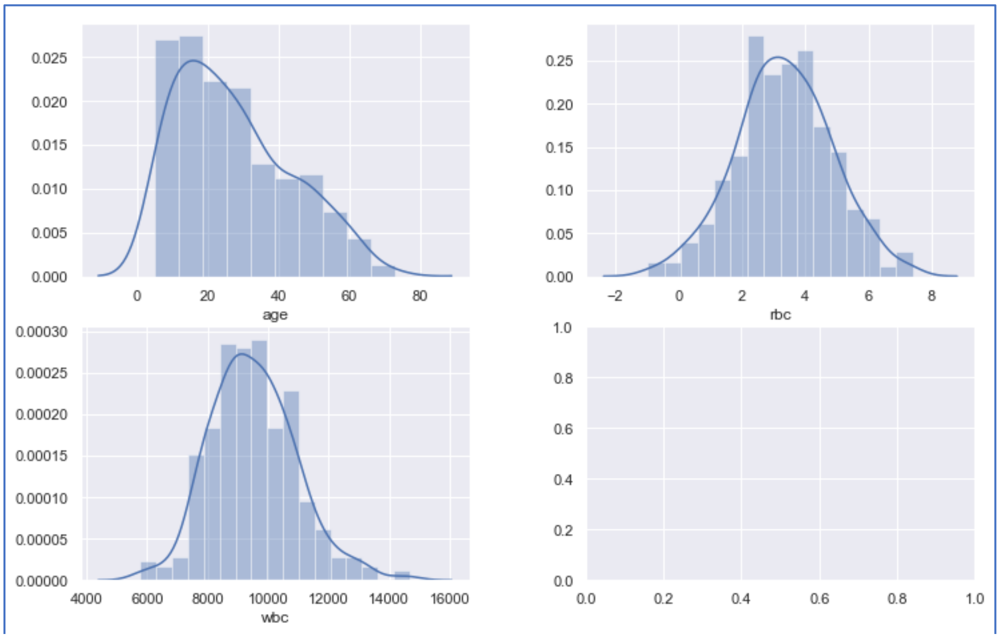
Conclusions:

A, Given a patient previously carrying mono/HIV, the odds of death for a patient previously had mono/HIV to the odds of death for a patient previously doesn’t carry mono/HIV is 0.2278. Under 95% confidence interval, the range of the odds ratio of death due to previous mono/HIV to without mono/HIV is between 0.081749 and 0.634867. Or in other words, we are expected to see about 77% decrease in odds of having Hodgkin’s disease from having mono/HIV to without mono/HIV.

B. An individual is expecting to experience 80.7% decrease in survival rate in odds from stage IV to stage I. In other words, a person is 80.7% more likely to survive from cancer stage I to cancer stage IV. So it is very likely that this person is not going to survive once in the cancer stage IV.

C. A 33-year old male with a red blood cell count of 4.3 cells/ul, a while blood cell count of 12,000 cells/mm3, no history of mono and no indication of HIV has been diagnosed with stage II Hodgkins has 71% survival rate with 69.2% to 75.45% (95% confidence interval).

Steps:
I first check the histogram of the sample data. Want to check whether the datasets follow normal distributions. Since we only have three potential factors: age, rbc and wbc. It turns out all of them are normally distributed.
Then I transferred all of the categorical predictor into numerical ones via one hot encoding.



Then I used the logistic regression model to model the relationship between feature matrix and response matrix.

From the logistic regression, I can get all of the key coefficients for the key of interested predictors such as the mono/HIV, stage IV, stage I and etc.

Model Coefficients:

Model Summary-Coefficients

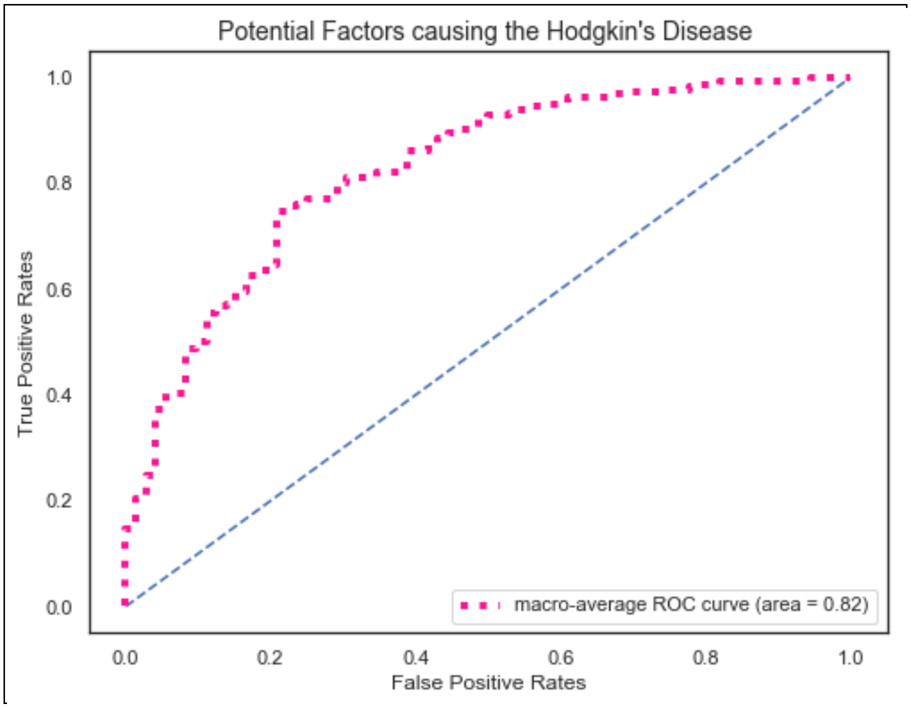
Factor	Coef	Std Error	Z	P>|z|	0.025	0.975
Age	-0.0321	0.01	-3.203	0.001	-0.052	-0.012
rbc	0.7392	0.114	6.469	0	0.515	0.963
wbc	3.84E-05	5.24E-05	0.733	0.463	-6.43E-05	0
StageII	0.5741	0.448	1.28	0.2	-0.305	1.453
StageIII	-0.1839	0.425	-0.433	0.665	-1.017	0.649
StageIV	-1.6445	0.446	-3.687	0	-2.519	-0.77
Gender_M	-0.0088	0.312	-0.028	0.977	-0.621	0.603
HIV.mon_Y	-1.4792	0.523	-2.829	0.005	-2.504	-0.454

Model Summary-Odds Ratio

Predictors	Odds Ratio	Lower CI	Upper CI
Age	1.000038	0.999936	1.000141
rbc	1.775467	0.737382	4.274968
wbc	0.832011	0.361738	1.913654
StageII	0.193116	0.080563	0.462917
StageIII	0.991206	0.537447	1.828066
StageIV	0.227815	0.081749	0.634867
Gender_M	1.000038	0.999936	1.000141
HIV.mon_Y	1.775467	0.737382	4.274968

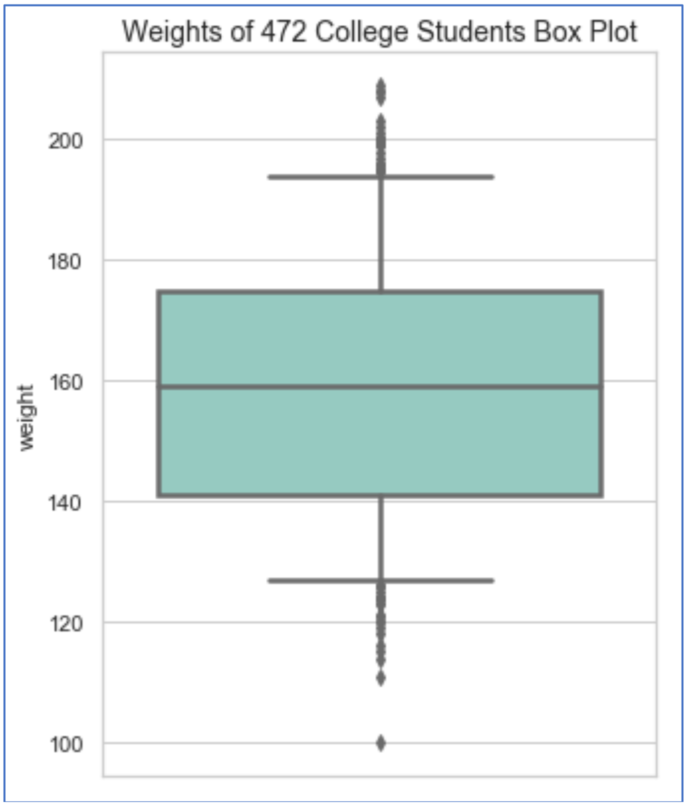
Model Performance.

I performed the ROC and calculated the AUC =0.82. Therefore, my logistic regression model has high sensitivity and specificity.
Next step is to perform the odds ratio based on the model parameters to quantify the mono/HIV's influence on death rates, and survival rates will be significantly improved comparing stage IV from stage I. As for the case study, we just need to run the patient's feature matrix via the model.



4.
The picture is the box plot for sample data 'weight.csv'.

As show in the picture, between each two lines extending vertically from the boxes outside the upper and lower quartiles and upper quartile and lower quartile are the whisker areas.
Outliers: Any data not included between the whiskers are plotted as the outliers
The upper whisker area is larger than lower whisker area.

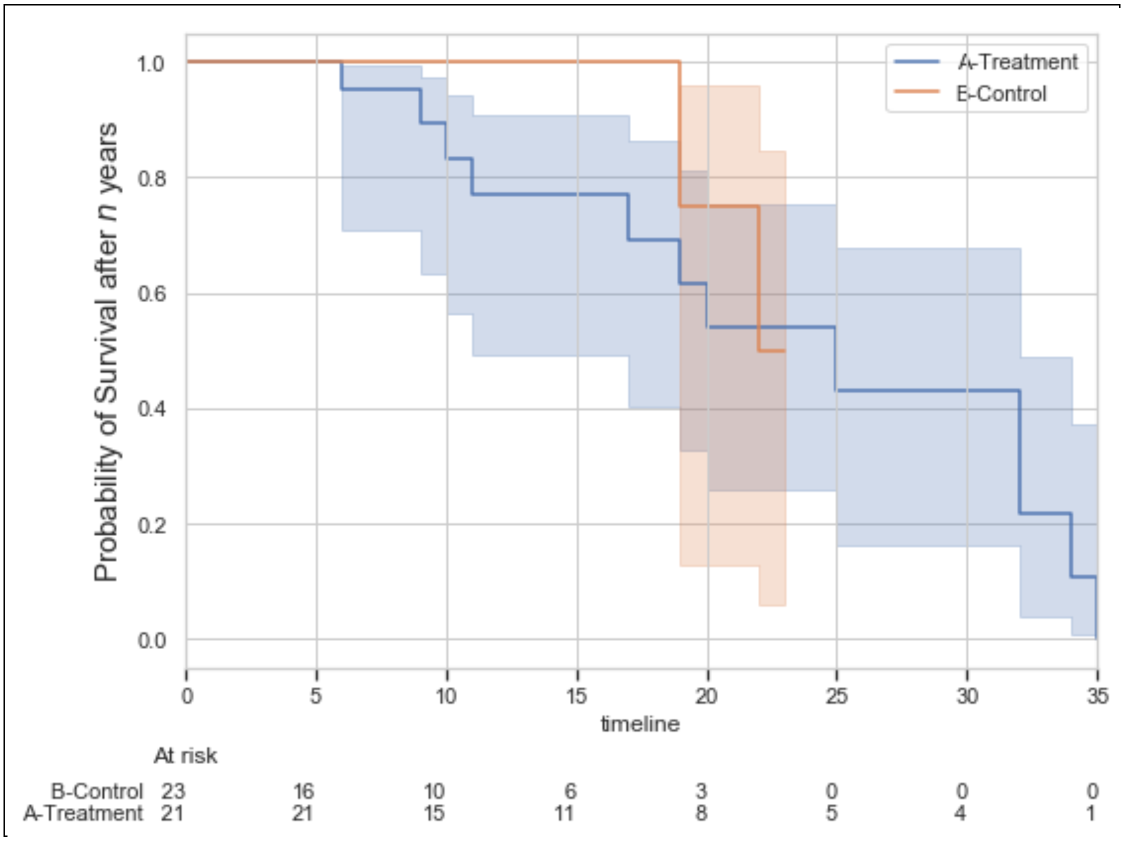


5, Conclusion: there are benefits to being placed in the treatment group as opposed to the control group due to the big difference of the number of uncensored persons between control and treatment groups.

I segment on ‘Treatment Group-A’ and ‘Control group-B’. It is incredible how higher survival rate for control group over treatment group. Control group does has a bias towards survival. The median of treatment group is also bigger than control group but difference is apparent in tails: if you are inside the treatment group and you have made past 25 years, you probably have a long life ahead. Meanwhile, for person in the control group, after 20 years, we don't know your probability of alive any more.

Data Prep:
1, First transfer categorical data into numerical one with one-hot encoding:
Result: ‘Death’->1, ‘Alive’->0
2, I used the KaplanMeierFitter metod fit the data. In our sample data, we fit the ‘Time’ and ‘Results’.
3, Using the ‘Survival_function_’ from the KaplanMeierFitter metod we got the survival analysis plot .

Median: Control VS Treatment	
Control Group	22
Treatment Group	25



The y-axis represents the probability a patient will be dead after t years, where t years is the on the x-axis. We see that for treatment group, once past 23 years, it is highly that this person is gonna still alive, however, once past 25 years the people in the control group is more likely to maintain a 0.43 death rate. Of course, like all good stats, we need to report how uncertain we are about these points estimate i.e. we need confidence intervals. As shown in the shadow areas in the plot.