CS 4372
ASSIGNMENT 1

Names of students in your group: Natalie Pedigo (nbp220000), Kylie Quinney (krq210000)

Number of free late days used: 0

Note: You are allowed a **total** of 4 free late days for the **entire semester**. You can use at most 2 for each assignment. After that, there will be a penalty of 10% for each late day.

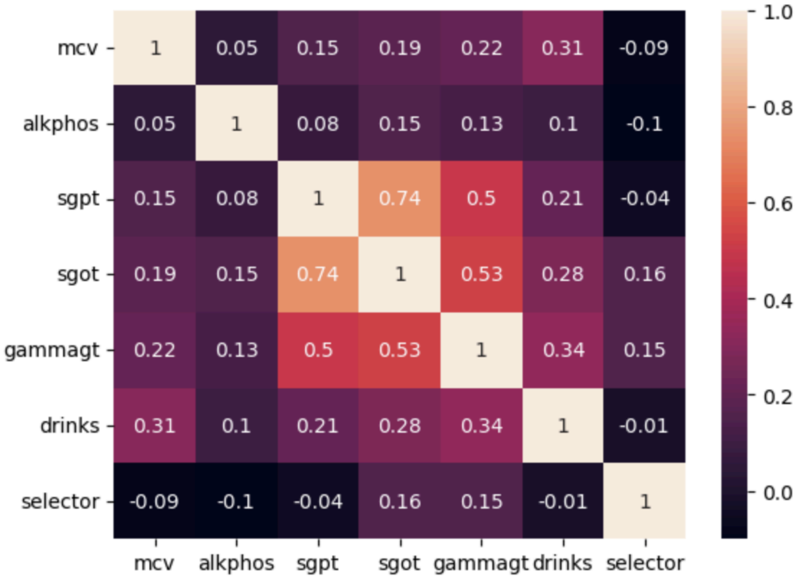
Overview

Our data contains patient blood test values which are thought to be sensitive to liver disorders that may arise due to alcohol consumption. We attempt to predict how many alcoholic beverages a person is consuming per day based on their blood work, using SGDRegressor and OLS methods.

Part One (SGDResgressor)

Plots

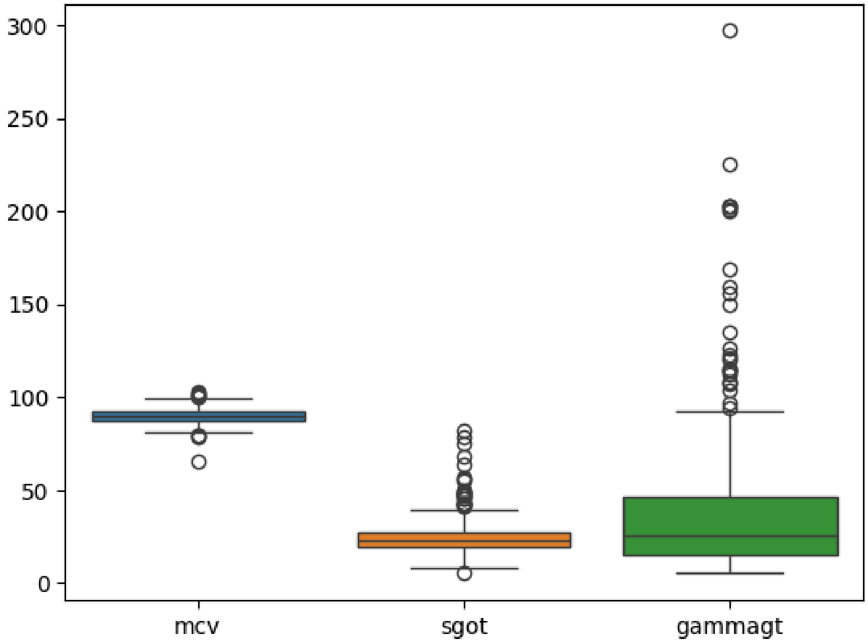
1. **Correlation Matrix**



	mcv	alkphos	sgpt	sgot	gammagt	drinks	selector
mcv	1	0.05	0.15	0.19	0.22	0.31	-0.09
alkphos	0.05	1	0.08	0.15	0.13	0.1	-0.1
sgpt	0.15	0.08	1	0.74	0.5	0.21	-0.04
sgot	0.19	0.15	0.74	1	0.53	0.28	0.16
gammagt	0.22	0.13	0.5	0.53	1	0.34	0.15
drinks	0.31	0.1	0.21	0.28	0.34	1	-0.01
selector	-0.09	-0.1	-0.04	0.16	0.15	-0.01	1

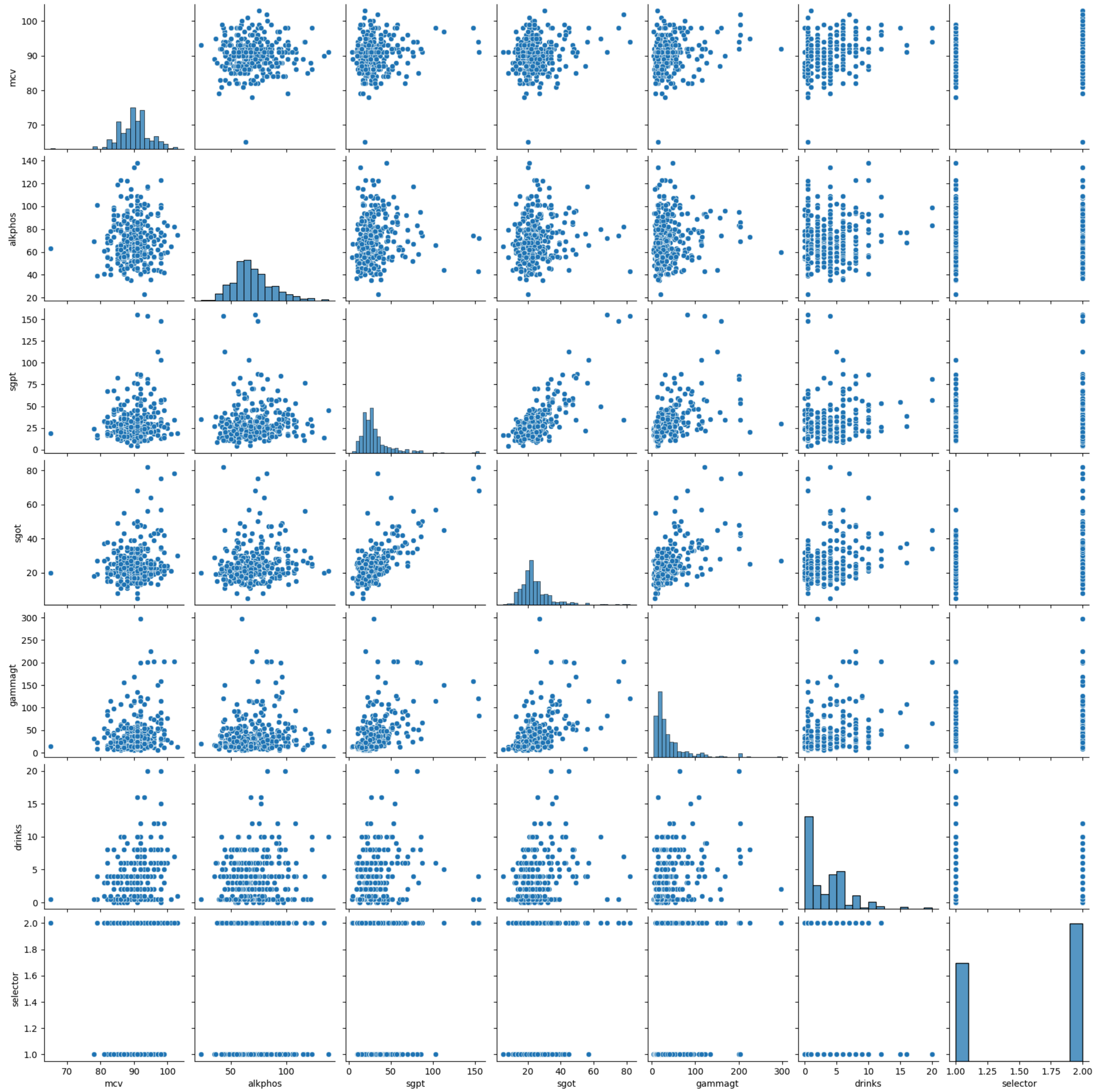
We used the heatmap correlation matrix to help us determine which features would be most useful in our model. We did this by choosing the features with the highest absolute value correlation with the target variable.

2. Box and Whisker Plot



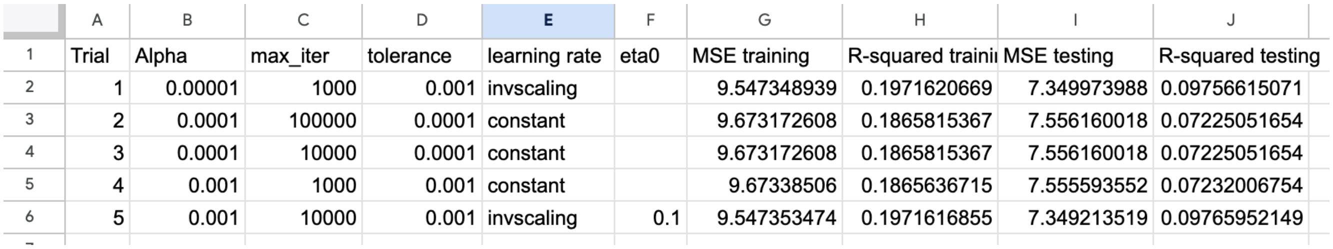
The above box and whisker plot shows us the distribution of our chosen features. From this plot we can see that each of these features has several outliers. We can also see that ‘gammagt’ in particular has a high volume of outliers and that its outliers may stray from the median by a large margin.

3. Pairplot Diagram



The above graphs are plots of every pair of variables, helping us visualize their relationships. These plots helped us choose our features when building the model. In the graphs where ‘drinks’ were plotted, we can see what kind of relationship it has, if any, with the independent variables.

Hyper-Parameter Experiment Log



	A	B	C	D	E	F	G	H	I	J
1	Trial	Alpha	max_iter	tolerance	learning rate	eta0	MSE training	R-squared trainii	MSE testing	R-squared testing
2	1	0.00001	1000	0.001	invscaling		9.547348939	0.1971620669	7.349973988	0.09756615071
3	2	0.0001	100000	0.0001	constant		9.673172608	0.1865815367	7.556160018	0.07225051654
4	3	0.0001	10000	0.0001	constant		9.673172608	0.1865815367	7.556160018	0.07225051654
5	4	0.001	1000	0.001	constant		9.67338506	0.1865636715	7.555593552	0.07232006754
6	5	0.001	10000	0.001	invscaling	0.1	9.547353474	0.1971616855	7.349213519	0.09765952149

Model Results

MSE Train: 9.547348939359688
MAE Train: 2.339358298250774
EV Train: 0.1971625190569012
R2 Train: 0.19716206688307059
Coefficients Train: [0.7732831, 0.24685804, 0.91934473]

MSE Test: 7.3499739880117065
MAE Test: 2.3650460852471693
EV Test: 0.10100685046118296
R2 Test: 0.09756615071180397
Coefficients Test: [0.7732831, 0.24685804, 0.91934473]

Analysis of SGDRegressor Model Results

For our linear regression model, we outputted the following statistics to better understand how our model performs on both training and testing data: MSE, MAE, EV, R^2, and Coefficients.

Our model produced a high MSE for both training and testing. This suggests that our model's predicted values varied widely from the target values. It is possible that this was caused by a high volume of outliers in the predictions of our model, due to the fact that MSE takes the square of the error, it is sensitive to larger numbers meaning it is more sensitive to outliers. We would not be able to come to this conclusion by looking at the MSE alone, but because the MAE is about 2.3 for both training and testing predictions, it does seem that outliers played a role in our high MSE.

Another observation is that our MSE for our training data was higher than that of our testing data by 0.002%. A lower MSE for testing predictions can suggest underfitting, but because the percent difference of ours is so low, this may not be the case. It could simply be that the data that happened to be used for testing does not represent the distribution well enough to be comparable, or it could be 'random' in the sense that if we were to plug in a different random state, that percentage difference may change.

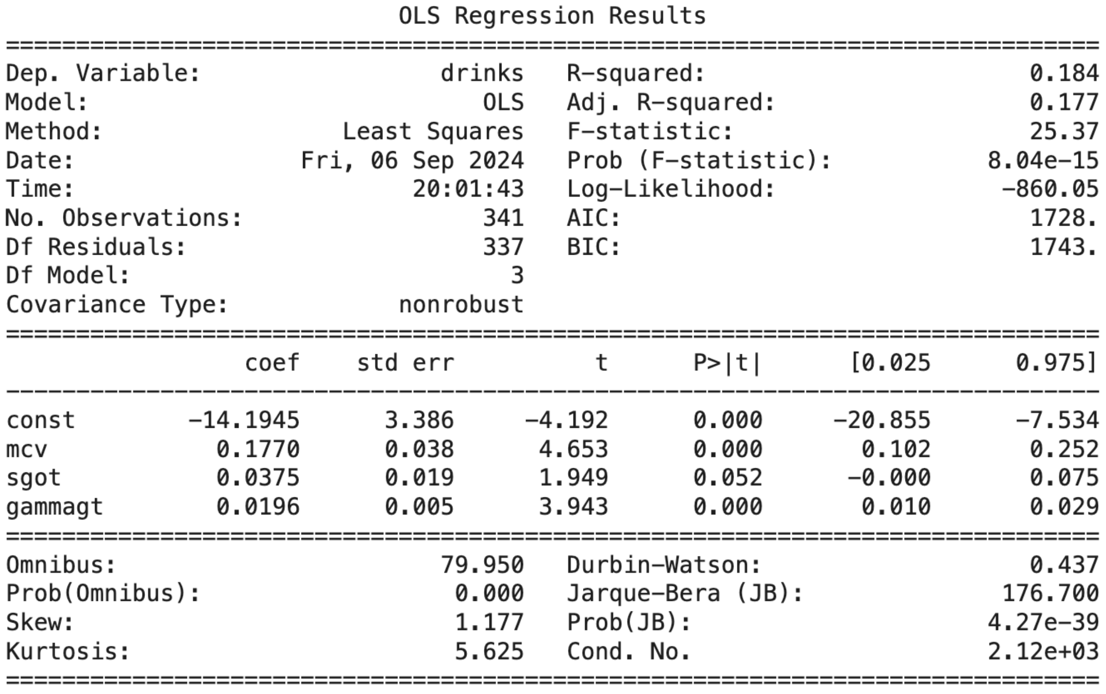
For our EV score, we got a 0.19 and 0.10 for training and testing data, respectively. The EV shows how well the predictions made by the model explains the variance in the target data, with a score close to 1 being ideal. Both of our EV values were far from 1, however, our training EV was nearly equal to our training R2, which means that our predictions were unbiased. We received a similar result for the EV and R2 of the testing data.

Our R2 themselves were also far lower than the preferred value of 1. However, due to the fact that the R2 of the OLS solution was a 0.18, it seems that the data we chose is not linear, meaning the independent variables (features) are not linearly correlated with the dependent variable (target), and therefore not an ideal candidate for a linear regression model.

For the coefficients, the weights are listed in the following order: mcv, sgot, gammagt. According to the heat map, we would have predicted that the coefficients from highest to lowest would be gammagt, mcv, sgot. However, it seems that our model weighted gammagt and mcv higher than sgot. However, the OLS solution did weight the features in the order that we predicted. Our model may have given the features different weights than we predicted for any number of reasons, such as outliers in the data, interactions between the features, or a non-linear relationship between the features and the target variable.

<u>Part Two (OLS)</u>

Regression Results



```
                            OLS Regression Results                            
==============================================================================
Dep. Variable:                 drinks   R-squared:                       0.184
Model:                            OLS   Adj. R-squared:                  0.177
Method:                 Least Squares   F-statistic:                     25.37
Date:                Fri, 06 Sep 2024   Prob (F-statistic):           8.04e-15
Time:                        20:01:43   Log-Likelihood:                -860.05
No. Observations:                 341   AIC:                             1728.
Df Residuals:                     337   BIC:                             1743.
Df Model:                           3                                         
Covariance Type:            nonrobust                                         
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const        -14.1945      3.386     -4.192      0.000     -20.855      -7.534
mcv            0.1770      0.038      4.653      0.000       0.102       0.252
sgot           0.0375      0.019      1.949      0.052      -0.000       0.075
gammagt        0.0196      0.005      3.943      0.000       0.010       0.029
==============================================================================
Omnibus:                       79.950   Durbin-Watson:                   0.437
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              176.700
Skew:                           1.177   Prob(JB):                     4.27e-39
Kurtosis:                       5.625   Cond. No.                     2.12e+03
==============================================================================
```

Analysis of OLS Model Results

Our dependent variable is the number of drinks, dependent on our features 'mcv', 'sgot', and 'gamma gt'. Our const coefficient, -14.1945, is the mean number of drinks predicted when our independent variables are equal to zero. As it is impossible for a person to consume negative drinks, this value is not very telling. Therefore, this value suggests that there is not a meaningful baseline of the dependent variable (drinks). The standard error of 3.386 shows that the variability of this estimate is relatively small compared to the value of our intercept (14.1945). The t-statistic, -4.192, and a p-value of 0 serve as strong evidence that the intercept estimate is statistically significant.

For our independent variable ‘mcv’ (average size of red blood cells in the sample), we calculated a coefficient of 0.177 and standard error of 0.038. Our coefficient indicates a positive relationship with our target, and our standard error being on the smaller side suggests our coefficient is precise and not likely to vary. The t-statistic, 4.653, is large enough to indicate that the coefficient is significantly different from zero. Our p-value of 0 suggests that there is enough evidence to confirm the variable has a significant impact on the target.

For the independent variable ‘sgot’ (level of SGOT enzyme in the blood), we calculated a coefficient of 0.0375 and a standard error of 0.019. While the coefficient shows a positive relationship, it is relatively small, suggesting that ‘sgot’ levels have minimal effect on the target. Although the standard error isn't very large, it's roughly half the size of the coefficient, thus indicating some uncertainty in the coefficient estimate. Our t-statistic of 1.949 is not very large, also indicating that our coefficient is only slightly significant. Our p-value of 0.052 is also just above the common threshold of 0.05, so we conclude there is not enough evidence to say this variable’s coefficient is significantly different from zero.

For the final independent variable ‘gammagt’ (level of GGT in the blood), we calculated a coefficient of 0.0196 and a standard error of 0.05. The coefficient is rather small, suggesting that although there is a positive relationship between this variable and the target, the impact is likely minimal. Our standard error is quite small compared to the coefficient value, suggesting a precise estimate of the coefficient with little variability. The t-statistic of 3.943 is large enough to suggest the coefficient is statistically significant. Also, with a p-value of 0, this indicates there is strong evidence that there is significant impact of GGT levels in the blood on the target, ‘drinks’.

With an R-squared value of just 0.184, we conclude that this linear model does not account for much of the variability in the target variable ‘drinks’. A possible reason for this could be the relationship between the variables are not linear, as we can see in our pair plots (plot 3). Another possibility is that important independent variables could be missing from this model. Whatever the reason, we are inclined to believe that this data is not linear. The adjusted R-squared value of 0.177 takes into account the number of independent variables used. Because our adjusted value is lower than our original R-squared, this suggests that there could be a predictor that does not meaningfully contribute to the model.

The last statistic we will discuss is our F-statistic, 25.37. This value is relatively high. Considering both our F-statistic and R-squared value together, we can conclude that the model is statistically significant, but our independent variables are rather weak. We can say that at least one predictor is contributing to explaining the variance in the target, however most of the variance remains unexplained.

Citations

Liver Disorders [Dataset]. (2016). UCI Machine Learning Repository. https://doi.org/10.24432/C54G67.

Sgdregressor. scikit. (n.d.). https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.SGDRegressor.html

GridSearchCV. scikit. (n.d.-a). https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html

GeeksforGeeks. (2024, August 7). *Interpreting the results of linear regression using OLS summary*. https://www.geeksforgeeks.org/interpreting-the-results-of-linear-regression-using-ols-summary/