Phyton Project Introduction to Health Informatics
Diana Ríos
202402000

DAILY ALCOHOL CONSUMPTION AND LIVER DISORDERS

According to the world health organization WHO, 400 million people (7% of the world population) aged 15 years or older have alcohol use disorders, and of them 209 million people (3.7% of the world population adult) live with alcohol dependence. Alcohol consumption is associated with risks of developing liver diseases, being an established carcinogen for the liver and other organs[1].

Since alcoholic toxicity is a topic of interest to me, I chose the database of Liver Disorders[2] (donated on 5/14/1990) for the development of this project. This data set has 5 variables which are all blood tests, that are biomarkers for liver disorders that might arise from excessive alcohol consumption.

** The code that I used, I put on a blue table*

```
!pip install pandas
!pip install matplotlib
!pip install seaborn
!pip install statsmodels
!pip install numpy
import pandas as pd
# Upload the data base "bupa.data" to the work environment
bupa = pd.read_csv('bupa.data', header=None)
bupa.columns = ['mcv', 'alkphos', 'sgpt', 'sgot', 'gammagt', 'drinks', 'selector']
bupa = pd.DataFrame(bupa)
bupa = bupa.drop('selector', axis=1)
print(bupa)
bupa = pd.DataFrame(bupa)
summary = bupa.describe()
print(summary)
```

This data set has [345 rows x 6 columns]

	mcv	alkphos	sgpt	sgot	gammagt	drinks
count	345.000000	345.000000	345.000000	345.000000	345.000000	345.000000
mean	90.159420	69.869565	30.405797	24.643478	38.284058	3.455072
std	4.448096	18.347670	19.512309	10.064494	39.254616	3.337835
min	65.000000	23.000000	4.000000	5.000000	5.000000	0.000000
25%	87.000000	57.000000	19.000000	19.000000	15.000000	0.500000
50%	90.000000	67.000000	26.000000	23.000000	25.000000	3.000000
75%	93.000000	80.000000	34.000000	27.000000	46.000000	6.000000
max	103.000000	138.000000	155.000000	82.000000	297.000000	20.000000

[1] World Health Organization. Topics: Alcohol. June 2024, available in: https://www.who.int/news-room/fact-sheets/detail/alcohol
[2] Liver Disorders [Dataset]. (2016). UCI Machine Learning Repository. https://doi.org/10.24432/C54G67

Cleaning the data set

In the data set description, they say that there are no missing values, but there is a note about duplicates:

```
Thanks to Leon for mentioning that there are duplicates in this data set.
--UCI ML Librarian

row 84 and 86:   94,58,21,18,26,2.0,2
row 141 and 318:   92,80,10,26,20,6.0,1
row 143 and 150:   91,63,25,26,15,6.0,1
row 170 and 176:   97,71,29,22,52,8.0,1
```

So, the duplicates were removed:

```
bupa.drop_duplicates(inplace=True)
print("Shape after removing duplicates:", bupa.shape)
print(bupa)
bupa = pd.DataFrame(bupa)
summary = bupa.describe()
print(summary)
```

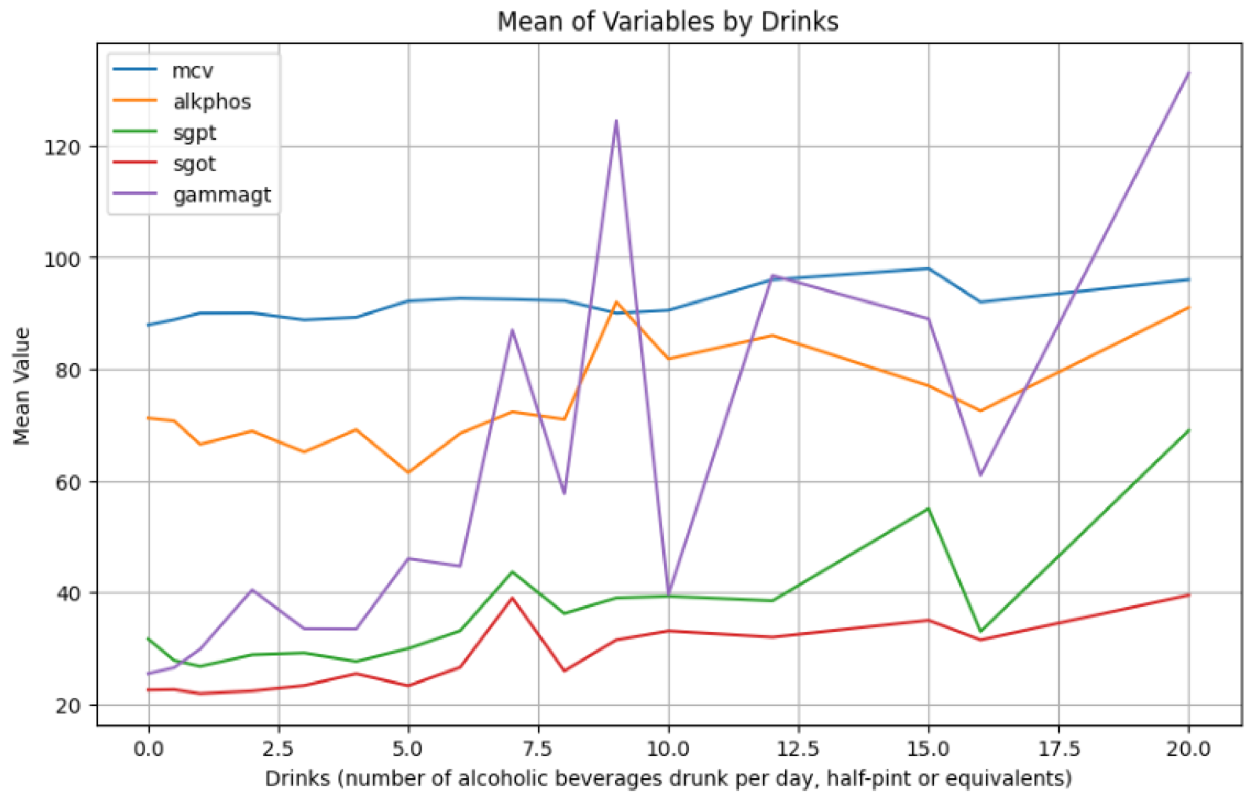
```
     mcv  alkphos  sgpt  sgot  gammagt  drinks
0     85       92    45    27       31     0.0
1     85       64    59    32       23     0.0
2     86       54    33    16       54     0.0
3     91       78    34    24       36     0.0
4     87       70    12    28       10     0.0
..   ...      ...   ...   ...      ...     ...
340   99       75    26    24       41    12.0
341   96       69    53    43      203    12.0
342   98       77    55    35       89    15.0
343   91       68    27    26       14    16.0
344   98       99    57    45       65    20.0

[341 rows x 6 columns]
              mcv     alkphos        sgpt        sgot     gammagt      drinks
count  341.000000  341.000000  341.000000  341.000000  341.000000  341.000000
mean    90.120235   69.891496   30.513196   24.662757   38.401760    3.431085
std      4.452385   18.431988   19.586249   10.115541   39.439379    3.341640
min     65.000000   23.000000    4.000000    5.000000    5.000000    0.000000
25%     87.000000   57.000000   19.000000   19.000000   15.000000    0.500000
50%     90.000000   67.000000   26.000000   23.000000   25.000000    3.000000
75%     92.000000   80.000000   34.000000   27.000000   46.000000    5.000000
max    103.000000  138.000000  155.000000   82.000000  297.000000   20.000000
```

Data Analysis

```
means = bupa.groupby('drinks')[['mcv', 'alkphos', 'sgpt', 'sgot', 'gammagt']].mean()

import matplotlib.pyplot as plt

plt.figure(figsize=(10, 6))

for column in means.columns:

    plt.plot(means.index, means[column], label=column)

plt.xlabel('Drinks (number of alcoholic beverages drunk per day, half-pint or equivalents)')

plt.ylabel('Mean Value')

plt.title('Mean of Biomarkers by Drinks')

plt.legend()

plt.grid(True)

print(

    "mcv: Mean Corpuscular Volume (fL)\n"
```

```
    "alkphos: Alkaline Phosphatase (U/L)\n"
    "sgpt: Alanine Transaminase (U/L)\n"
    "sgot: Aspartate Transaminase (U/L)\n"
    "gammagt: Gamma-Glutamyl Transpeptidase (U/L)"
)
plt.show()
```

```
mcv: Mean Corpuscular Volume (fL)
alkphos: Alkaline Phosphatase (U/L)
sgpt: alanine transaminase (U/L)
sgot: aspartate transaminase (U/L)
gammagt: gamma-glutamyl transpeptidase (U/L)
```



This first plot was made to show the change of the mean value of the different biomarkers according to the number of beverages per day. We can see a tendency to increase blood parameters when the number of alcoholic drinks per day increases. However, this graph must be analyzed with caution, since for some averages, only one patient's data was available. Likewise, it is not possible to know what is considered risky alcohol consumption.

According to the National Institute on Alcohol Abuse and Alcoholism of the USA[3], there are different terms that describe different patterns of alcohol consumption, to evaluate and make informed decisions, and they are different in females and males. In this data set, the variable sex is not included but trying to have an approach

[3] National Institute on Alcohol Abuse and Alcoholism, Understanding Alcohol Drinking Patterns. January 2025. Available in: https://www.niaaa.nih.gov/alcohols-effects-health/alcohol-drinking-patterns

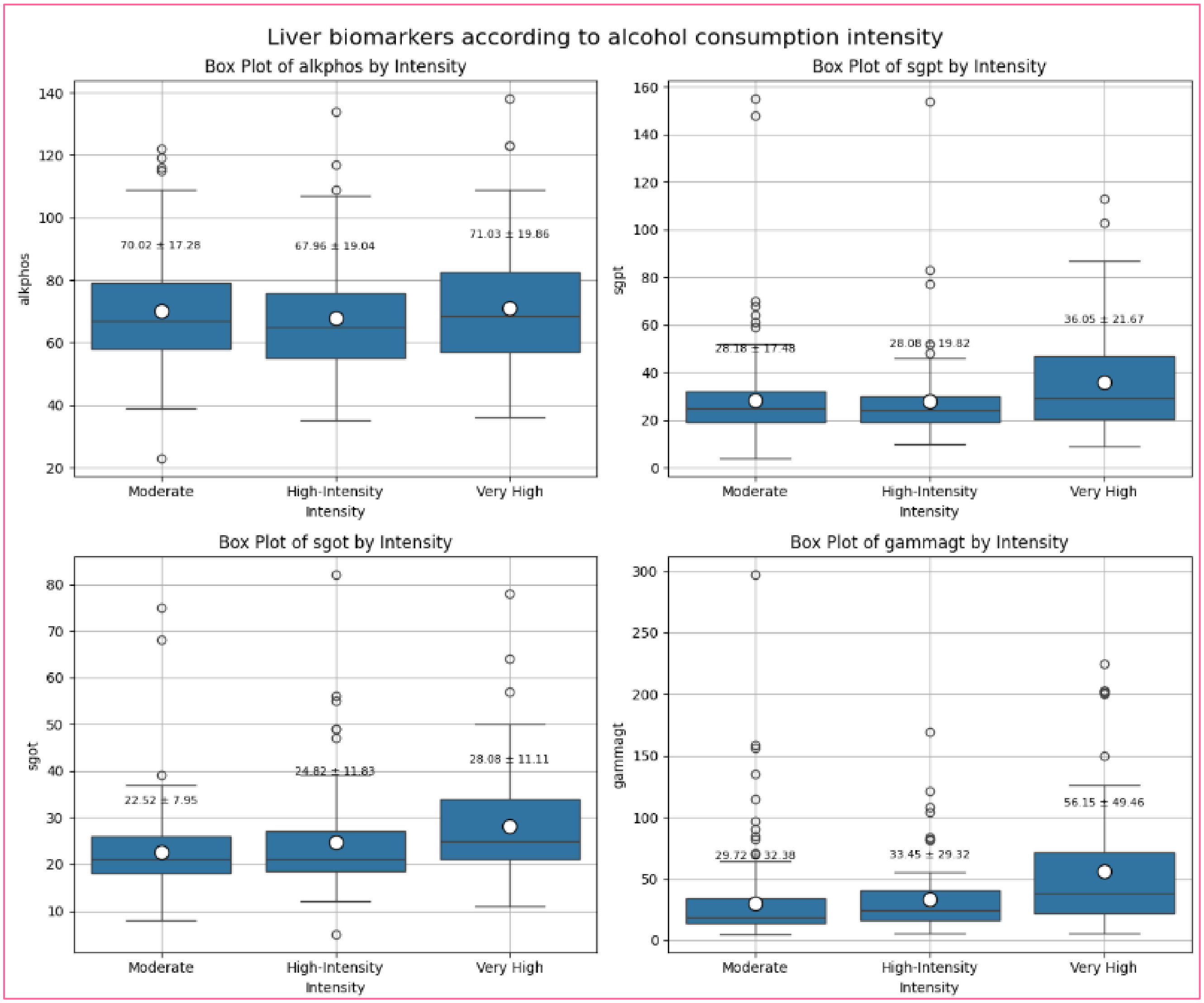
to the proposed consumption patterns, it was decided to divide the dataset like this: Drinking in Moderation (2 drinks maximum), High-Intensity Drinking (Between 3 and 4 drinks) and Heavy Drinking (5 drinks or more).

```
def Intensity(valor):
    if 0 <= valor <= 2: # Use 'valor' instead of 'drinks'
        return 'Moderate'
    elif 3 <= valor <= 4: # Use 'valor' instead of 'drinks'
        return 'High-Intensity'
    elif valor >= 5: # Use 'valor' instead of 'drinks'
        return 'Very High'
bupa['Intensity'] = bupa['drinks'].apply(Intensity)

intensity_summary = bupa.groupby('Intensity').describe()
print(intensity_summary)
import seaborn as sns

frequency_table = bupa['Intensity'].value_counts().reset_index()
frequency_table.columns = ['Intensity', 'Frequency']
frequency_table['Percent'] = (frequency_table['Frequency'] / frequency_table['Frequency'].sum()) * 100
print(frequency_table)
```

With this classification and the grouping carried out we can see that most cases correspond to moderate drinkers (49.27%) while High-Intensity consumption is 20.82%).

	Intensity	Frequency	Percent
0	Moderate	168	49.266862
1	Very High	102	29.912023
2	High-Intensity	71	20.821114

```
import numpy as np

columns_to_plot = ['alkphos', 'sgpt', 'sgot', 'gammagt']

fig, axes = plt.subplots(2, 2, figsize=(12, 10))
axes = axes.flatten()

for i, column in enumerate(columns_to_plot):
    ax = axes[i]
    sns.boxplot(x='Intensity', y=column, data=bupa, showmeans=True,
            meanprops={"marker": "o", "markerfacecolor": "white",
                    "markeredgecolor": "black", "markersize": "10"}, ax=ax)


    for intensity in bupa['Intensity'].unique():
        mean = bupa[bupa['Intensity'] == intensity][column].mean()
        std = bupa[bupa['Intensity'] == intensity][column].std()
        ax.text(bupa['Intensity'].unique().tolist().index(intensity),
            mean + std + 2,
            f'{mean:.2f} ± {std:.2f}', ha='center', va='bottom', fontsize=8)

    ax.set_title(f'Box Plot of {column} by Intensity')
    ax.set_ylabel(column)
    ax.set_xlabel('Intensity')
    ax.grid(True)

plt.suptitle('Liver biomarkers according to alcohol consumption intensity', fontsize=16)
plt.tight_layout()
plt.show()
```



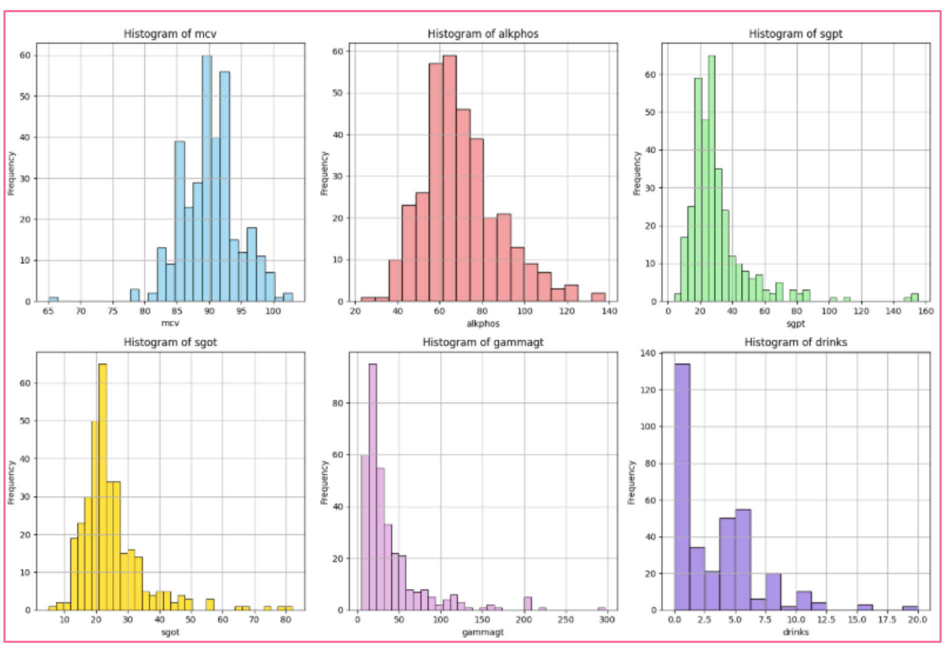
In the boxplot we can see a difference in the means in some variables, so it was decided to run an Anova test for each of them, although the standard deviations were very high given the difference between the data.

```
import statsmodels.formula.api as smf
import statsmodels.api as sm

model = smf.ols('alkphos ~ Intensity', data=bupa).fit()
anova_table = sm.stats.anova_lm(model, typ=2)

print("ANOVA Table:")
print(anova_table.iloc[:1])
model = smf.ols('sgpt ~ Intensity', data=bupa).fit()
anova_table = sm.stats.anova_lm(model, typ=2)
print("ANOVA Table:")
print(anova_table.iloc[:1])

model = smf.ols('sgot ~ Intensity', data=bupa).fit()
anova_table = sm.stats.anova_lm(model, typ=2)
print("ANOVA Table:")
print(anova_table.iloc[:1])

model = smf.ols('gammagt ~ Intensity', data=bupa).fit()
anova_table = sm.stats.anova_lm(model, typ=2)
print("ANOVA Table:")
print(anova_table.iloc[:1])
```

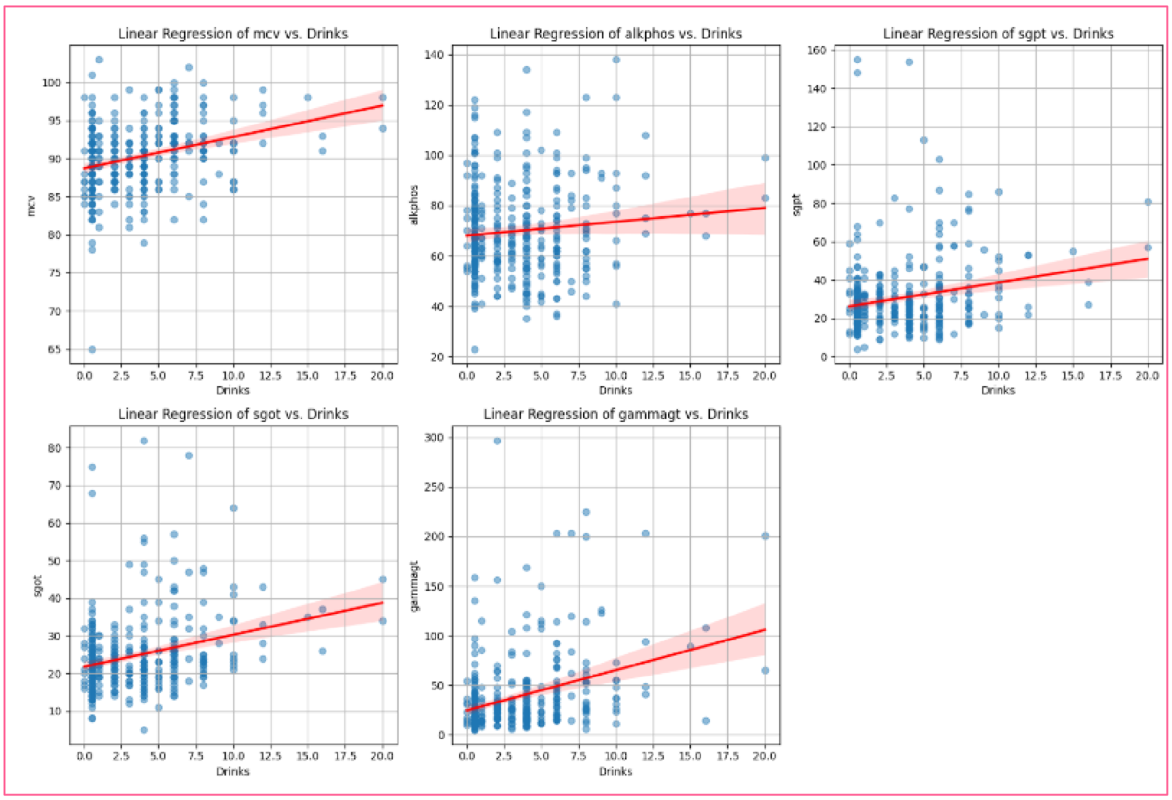
Variable	Anova	Interpretation
alkphos	ANOVA Table: sum_sq df F PR(>F) Intensity 400.253905 2.0 0.587633 0.556207	There is not difference between means
sgpt	ANOVA Table: sum_sq df F PR(>F) Intensity 4460.299899 2.0 5.983848 0.002794	There is difference between means
sgot	ANOVA Table: sum_sq df F PR(>F) Intensity 1960.31998 2.0 10.091231 0.000055	There is difference between means
gammagt	ANOVA Table: sum_sq df F PR(>F) Intensity 46521.736171 2.0 16.300193 1.745142e-07	There is difference between means

Although some variables do not have an apparently normal distribution, as seen in the histograms, I decided to also perform a linear regression to make the corresponding graph, just for the purpose of the exercise in phyton:

```
import matplotlib.pyplot as plt
import statsmodels.formula.api as smf

columns_to_plot = ['mcv', 'alkphos', 'sgpt', 'sgot', 'gammagt', 'drinks']
colors = ['skyblue', 'lightcoral', 'lightgreen', 'gold', 'plum', 'mediumpurple']

fig, axes = plt.subplots(2, 3, figsize=(15, 10))
axes = axes.flatten()

for i, column in enumerate(columns_to_plot):
    ax = axes[i]
    sns.histplot(bupa[column], bins='auto', color=colors[i], edgecolor='black', ax=ax)
    ax.set_title(f'Histogram of {column}')
    ax.set_xlabel(column)
    ax.set_ylabel('Frequency')
    ax.grid(True)

plt.tight_layout()
plt.show()
```



```
columns_to_plot = ['mcv', 'alkphos', 'sgpt', 'sgot', 'gammagt']

fig, axes = plt.subplots(2, 3, figsize=(15, 10))
axes = axes.flatten()

for i, column in enumerate(columns_to_plot):
    ax = axes[i]
    sns.regplot(x='drinks', y=column, data=bupa, ax=ax,
                scatter_kws={'alpha': 0.5},
                line_kws={'color': 'red'})

    ax.set_title(f'Linear Regression of {column} vs. Drinks')
    ax.set_xlabel('Drinks')
    ax.set_ylabel(column)
    ax.grid(True)

for j in range(len(columns_to_plot), len(axes)):
    axes[j].set_visible(False)

plt.tight_layout()
plt.show()
```




```
columns_to_plot = ['mcv', 'alkphos', 'sgpt', 'sgot', 'gammagt']
results_summary = []

for column in columns_to_plot:
    model = smf.ols(formula=f'{column} ~ drinks', data=bupa).fit()
    results_summary.append([column, model.params['drinks'], model.pvalues['drinks'], model.rsquared])

results_df = pd.DataFrame(results_summary, columns=['Variable', 'Coefficient', 'P-value', 'R-squared'])

print(results_df)
```

	Variable	Coefficient	P-value	R-squared
0	mcv	0.412030	5.440444e-09	0.095629
1	alkphos	0.551265	6.527227e-02	0.009988
2	sgpt	1.236244	8.680910e-05	0.044486
3	sgot	0.849852	1.351856e-07	0.078818
4	gammagt	4.057925	6.764775e-11	0.118214

Again, with these p-values, alkaline phosphatase does not seem to have a relationship with the intensity of alcohol consumption. In contrast, the other liver variables seem to be sensitive markers for liver damage related to the degree of alcohol consumption, according to the different tests performed.

Discusion

Biological markers of alcohol consumption are those that indicate liver damage. The mean erythrocyte corpuscular volume (MCV) increases due to direct damage to the hematopoietic stem cell, but also due to vitamin B12 deficiency. Regarding liver enzymes, mainly gamma glutamyl transpeptidase (gammagt) and aspartate aminotransferase (sgot), their serum levels also increase in alcoholic patients, with aspartate aminotransferase (sgot) being higher than alanine aminotransferase (sgpt), with a ratio sgot/sgpt greater than two. The increase in gammagt occurs because alcohol is a powerful inducer of the hepatic microsomal system, the fundamental seat of this enzyme. And alkaline phosphatase (alkphos) usually increases in those who consume large amounts and in cholestasis[4][5].

The Mayo Clinic[6] indicates the following reference values for markers of liver damage:

- Alamine aminotransferase. 7 to 55 units per liter (U/L).
- Aspartate aminotransferase. 8 to 48 U/L.
- Alkaline phosphotase. 40 to 129 U/L.
- Gamma-glutamyl transpeptidase. 8 to 61 U/L.

With this database, we were able to find statistically significant differences between the means according to the intensity of alcohol consumption for the variables MCV, gammagt, sgot and sgpt, also finding a positive correlation in the linear regression. Alkaline phosphatase showed no statistical difference in the means according to consumption or correlation in linear regression. The mean value for Very High consumption did not exceed the reference values.

These data were analyzed for pedagogical purposes, but it should be noted that they are not normally distributed, and their standard deviations were very high.

[4] Ll Caballería, J. Caballería, A. Parés. Enfermedad hepática alcohólica. Medicina Integral. 2000,35:10.474-480

[5] M. Marcos Martín, I. Pastor Encinas y F. J. Laso Guzmán. Marcadores biológicos del alcoholismo. Rev Clin Esp. 2005;205(9):443-5

[6] Mayo Clinic. Liver function tests. 2025. Available in: https://www.mayoclinic.org/tests-procedures/liver-function-tests/about/pac-20394595