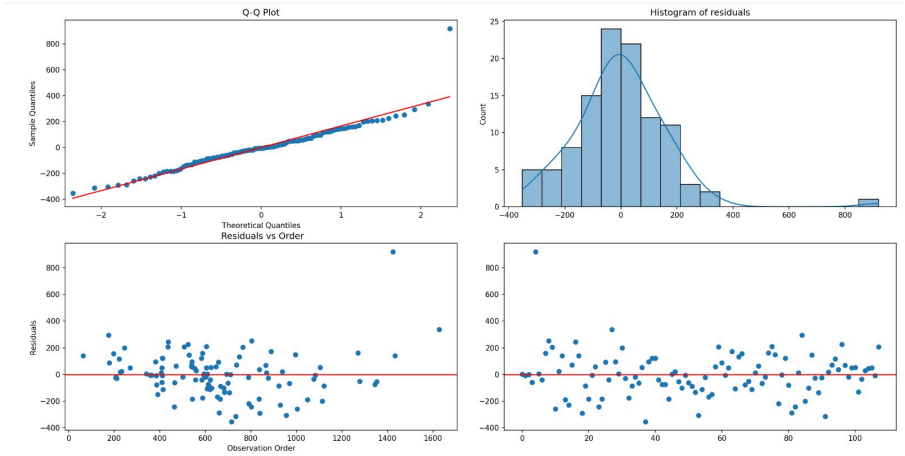
Regresión Lineal Múltiple

```python
# Load dataset
df = pd.read_csv('EjercicioEstadistica.csv')

# Modelo Base
X = df[['Factor_Coagulacion', 'Indice_Pronostico', 'Funcion_de_enzima', 'Funcion_de_higado', 'Edad', 'Genero', 'Alcohol_moderado', 'Alcohol_severo']]
y = df['Sobrevivencia_dias']

scaler = MinMaxScaler()
X_scaled = scaler.fit_transform(X)
X_scaled = sm.add_constant(X_scaled)

X_train, X_test, y_train, y_test = train_test_split(X_scaled, y, test_size=0.2, random_state=0)
model = sm.OLS(y_train, X_train).fit()

print(model.summary())

# Variables significativas
significant_vars = model.pvalues[model.pvalues < 0.05].index.tolist()
print("\nVariables significativas seleccionadas:")
print(significant_vars)

# Modelo Variables significativas
X_train_significant = pd.DataFrame(X_train, columns=model.params.index)[significant_vars]
X_test_significant = pd.DataFrame(X_test, columns=model.params.index)[significant_vars]

X_train_significant = X_train_significant.reset_index(drop=True)
y_train = y_train.reset_index(drop=True)
significant_model = sm.OLS(y_train, X_train_significant).fit()
print(significant_model.summary())

# Modelo Polynomial
formula = 'Sobrevivencia_dias ~ Factor_Coagulacion * Edad + I(Indice_Pronostico**2) + Funcion_de_enzima + Funcion_de_higado + Genero + Alcohol_moderado + Alcohol_severo'
poly_model = smf.ols(formula, data=df).fit()
print("\nModelo con Interacciones y Términos Polinomiales:")
print(poly_model.summary())

# Compare models
print("\nComparación de Modelos:")
print(f"R² del modelo base: {model.rsquared}")
print(f"R² ajustado del modelo base: {model.rsquared_adj}")

print(f"R² del modelo significativo: {significant_model.rsquared}")
print(f"R² ajustado del modelo significativo: {significant_model.rsquared_adj}")

print(f"R² del modelo polinómico: {poly_model.rsquared}")
print(f"R² ajustado del modelo polinómico: {poly_model.rsquared_adj}")

# ANOVA del mejor modelo
anova_table = sm.stats.anova_lm(poly_model, typ=2)
print("\nTabla de ANOVA:")
print(anova_table)
```

```python
# Verificacion de supuestos

residuals = poly_model.resid
fitted = poly_model.fittedvalues

fig, axs = plt.subplots(2,2, figsize=(12,10))
sm.qqplot(residuals,line='s',ax=axs[0,0])
axs[0,0].set_title('Q-Q Plot')

sns.histplot(residuals, kde=True, ax=axs[0,1])
axs[0,1].set_title('Histogram of residuals')

axs[1,0].scatter(fitted, residuals)
axs[1,0].axhline(y=0, color='r', linestyle='-')
axs[1,0].set_title(' Residuals vs Fitted')
axs[1,0].set_xlabel('Fitted values')
axs[1,0].set_ylabel('Residuals')

axs[1,1].scatter(range(len(residuals)),residuals)
axs[1,1].axhline(y=0, color='r', linestyle='-')
axs[1,0].set_title(' Residuals vs Order')
axs[1,0].set_xlabel('Observation Order')
axs[1,0].set_ylabel('Residuals')

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

	feature	VIF
0	Factor_Coagulacion	18.461696
1	Indice_Pronostico	12.991196
2	Funcion_de_enzima	14.622709
3	Funcion_de_higado	16.744367
4	Edad	14.977210
5	Genero	2.050575
6	Alcohol_moderado	2.763062
7	Alcohol_severo	1.791222



En el gráfico Q-Q podemos ver que la mayoría de los puntos se adhieren a la línea roja, mostrando una distribución normal. Aunque hay un valor atípico que se dispara al final, del lado derecho. En el histograma de residuos podemos observar lo mismo, con una distribución normal centrada en 0, con los valores atípicos del lado derecho. Tanto los residuos vs orden, y residuos vs variables predictoras presentan una distribución bastante aleatoria, por lo que no hay correlación.

Género, alcohol moderado, y alcohol severo tiene colinealidad moderada, mientras que edad , función de hígado, de enzima y factor de coagulación, tienen una colinealidad extremadamente alta.

Modelo Base:.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:     Sobrevivencia_dias   R-squared:                       0.792
Model:                            OLS   Adj. R-squared:                  0.771
Method:                 Least Squares   F-statistic:                     36.72
Date:                Mon, 09 Sep 2024   Prob (F-statistic):           2.94e-23
Time:                        22:17:26   Log-Likelihood:                -561.02
No. Observations:                  86   AIC:                             1140.
Df Residuals:                      77   BIC:                             1162.
Df Model:                           8
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const       -633.1217    117.881     -5.371      0.000    -867.852    -398.391
x1           512.3503    143.000      3.583      0.001     227.600     797.101
x2           776.6339    123.633      6.282      0.000     530.450    1022.818
x3           934.7404    117.788      7.936      0.000     700.195    1169.286
x4           408.2303    167.681      2.435      0.017      74.334     742.126
x5           -19.2630     67.539     -0.285      0.776    -153.750     115.224
x6             0.8462     38.358      0.022      0.982     -75.534      77.226
x7           -44.4929     44.943     -0.990      0.325    -133.985      44.999
x8           196.0206     56.249      3.485      0.001      84.016     308.026
==============================================================================
Omnibus:                       38.379   Durbin-Watson:                   1.820
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              160.286
Skew:                           1.293   Prob(JB):                     1.56e-35
Kurtosis:                       9.168   Cond. No.                         20.9
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

El modelo base tiene una R^2 de 0.792 el cual representa un buen modelo. Ya ajustado presenta 0.771, el cual sigue siendo un valor relativamente alto.

No significativas:
Las variables con valores p más altos fueron, genero, edad y alcohol moderado.

Significativas
Las variables con valores p más bajos fueron función de enzima, función de hígado y factor de coagulación.

Modelo Significativo:

```
Variables significativas seleccionadas:
['const', 'x1', 'x2', 'x3', 'x4', 'x8']
                            OLS Regression Results
==============================================================================
Dep. Variable:     Sobrevivencia_dias   R-squared:                       0.789
Model:                            OLS   Adj. R-squared:                  0.776
Method:                 Least Squares   F-statistic:                     59.98
Date:                Mon, 09 Sep 2024   Prob (F-statistic):          1.22e-25
Time:                        22:19:45   Log-Likelihood:                -561.62
No. Observations:                  86   AIC:                             1135.
Df Residuals:                      80   BIC:                             1150.
Df Model:                           5
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const       -667.0195    105.836     -6.302      0.000    -877.641    -456.398
x1           494.6147    140.187      3.528      0.001     215.634     773.596
x2           772.4119    120.584      6.406      0.000     532.442    1012.382
x3           931.7933    116.193      8.019      0.000     700.561    1163.026
x4           427.2664    164.546      2.597      0.011      99.809     754.724
x8           227.0977     46.493      4.885      0.000     134.574     319.622
==============================================================================
Omnibus:                       38.049   Durbin-Watson:                   1.752
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              149.763
Skew:                           1.307   Prob(JB):                     3.02e-33
Kurtosis:                       8.913   Cond. No.                         17.7
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

El modelo significativo tiene una R^2 de 0.789 el cual representa un buen modelo. Ya ajustado presenta 0.776, el cual sigue siendo un valor relativamente alto. Aunque en general peor que el modelo base.

No significativas
Las variables con valores p más altos fueron, función de hígado y factor de coagulación.

Significativas
Las variables con valores p más bajos fueron función de enzima, alcohol severo y factor de coagulación.

Modelo con Interacciones y Términos Polinomiales:

```
Modelo con Interacciones y Términos Polinomiales:
                            OLS Regression Results
==============================================================================
Dep. Variable:     Sobrevivencia_dias   R-squared:                       0.769
Model:                            OLS   Adj. R-squared:                  0.748
Method:                 Least Squares   F-statistic:                     36.33
Date:                Mon, 09 Sep 2024   Prob (F-statistic):           2.04e-27
Time:                        22:19:45   Log-Likelihood:                -705.51
No. Observations:                 108   AIC:                             1431.
Df Residuals:                      98   BIC:                             1458.
Df Model:                           9
Covariance Type:            nonrobust
================================================================================================
                                   coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------------------
Intercept                    -644.1234    351.082     -1.835      0.070   -1340.833      52.587
Factor_Coagulacion             21.1123     59.323      0.356      0.723     -96.612     138.837
Edad                           -2.6285      6.742     -0.390      0.697     -16.007      10.750
Factor_Coagulacion:Edad         0.5891      1.134      0.519      0.605      -1.662       2.840
I(Indice_Pronostico ** 2)       0.0642      0.009      6.823      0.000       0.046       0.083
Funcion_de_enzima               8.7261      1.036      8.422      0.000       6.670      10.782
Funcion_de_higado              78.6854     26.070      3.018      0.003      26.950     130.421
Genero                          7.3532     34.849      0.211      0.833     -61.803      76.509
Alcohol_moderado              -27.8982     40.232     -0.693      0.490    -107.738      51.942
Alcohol_severo                201.6155     50.997      3.954      0.000     100.415     302.816
==============================================================================
Omnibus:                       49.495   Durbin-Watson:                   1.779
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              261.198
Skew:                           1.381   Prob(JB):                     1.91e-57
Kurtosis:                      10.100   Cond. No.                     9.96e+04
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 9.96e+04. This might indicate that there are
strong multicollinearity or other numerical problems.
```

El modelo significativo tiene una R^2 de 0.769 el cual representa un buen modelo. Ya ajustado presenta 0.748, el cual sigue siendo un valor relativamente alto. Aunque en general peor que el modelo base e incluso que el modelo con variables significativas, siendo el peor hasta el momento.

No Significativas
Las variables con valores p más altos fueron, genero , factor de coagulación, edad y alcohol moderado.

Significativas
Las variables con valores p más bajos fueron función de enzima, función de hígado y alcohol severo.

Comparación de Modelos y Tabla ANOVA:

```
Comparación de Modelos:
R² del modelo base: 0.7923319698299689
R² ajustado del modelo base: 0.7707560705915241
R² del modelo significativo: 0.7894137156913739
R² ajustado del modelo significativo: 0.7762520729220848
R² del modelo polinómico: 0.7694001756866596
R² ajustado del modelo polinómico: 0.7482226408007406

Tabla de ANOVA:
                               sum_sq    df          F        PR(>F)
Factor_Coagulacion         3.592822e+05   1.0  11.793831  8.725071e-04
Edad                       8.700511e+03   1.0   0.285604  5.942619e-01
Factor_Coagulacion:Edad    8.218180e+03   1.0   0.269771  6.046567e-01
I(Indice_Pronostico ** 2)  1.418278e+06   1.0  46.556526  7.426172e-10
Funcion_de_enzima          2.161030e+06   1.0  70.938193  3.156798e-13
Funcion_de_higado          2.775068e+05   1.0   9.109465  3.240918e-03
Genero                     1.356319e+03   1.0   0.044523  8.333224e-01
Alcohol_moderado           1.464809e+04   1.0   0.480840  4.896825e-01
Alcohol_severo             4.761540e+05   1.0  15.630279  1.455129e-04
Residual                   2.985429e+06  98.0        NaN           NaN
```

El mejor modelo fue el base (R^2 de 0.792), el modelo polinómico ajustado dio los peores resultados (R^2 de 0.748), por lo que no mejora la capacidad de nuestro modelo para predecir los datos.

La edad, género y alcohol moderado no son indicadores significativos, mientras que el factor de coagulación, índice pronóstico cuadrado, función de enzima, así como el alcohol moderado, sí fueron altamente significativos.

El modelo base y el modelo significativo tienen presiones parecidas, siendo ambos los mayores en cuestión de exactitud, por lo que podemos decir que las variables seleccionadas por el significativo fueron las correctas.