

```
In [1]: import pandas as pd
import numpy as np
import matplotlib.pyplot as plt

df = pd.read_csv('e11.csv')

for i in range(1,df.shape[1]):
    df.iloc[:,i].replace(['#REF!', '#N/A!', ''], np.nan, inplace=True)
    df.iloc[:, i] = pd.to_numeric(df.iloc[:, i], errors='coerce')

threshold = 0.5
columns_to_drop_missing = df.columns[df.isna().mean() > threshold]
df_copy = df.copy()
df_copy.drop(columns=columns_to_drop_missing, inplace = True)
print(columns_to_drop_missing)
```

```
Index(['c188', 'c189', 'c190', 'c199', 'c202', 'c204', 'c206', 'c223', 'c226',
      'c229', 'c231', 'c232', 'c233', 'c234'],
      dtype='object')
```

```
<ipython-input-1-7432f26b8866>:9: DeprecationWarning: In a future version,
`df.iloc[:, i] = newvals` will attempt to set the values inplace instead of
always setting a new array. To retain the old behavior, use either `df[df.
columns[i]] = newvals` or, if columns are non-unique, `df.isetitem(i, new
vals)`
```

```
df.iloc[:, i] = pd.to_numeric(df.iloc[:, i], errors='coerce')
```

```
In [1]:
```

```

In [2]: from sklearn.linear_model import HuberRegressor
import numpy as np
# Assuming df is your DataFrame
columns_to_exclude = ['c51', 'c52', 'c53', 'c54', 'c1']

# Exclude specified columns
columns_to_process = [col for col in df_copy.columns if col not in columns_to_exclude]

# Convert columns to numeric, skipping the first row
df_copy[columns_to_process] = df_copy.iloc[1:][columns_to_process].apply(pd.to_numeric, errors='coerce')

# Calculate IQR for each column
Q1 = df_copy[columns_to_process].quantile(0.25)
Q3 = df_copy[columns_to_process].quantile(0.75)
IQR = Q3 - Q1

# Define lower and upper bounds
LB = Q1 - 1.5 * IQR
UB = Q3 + 1.5 * IQR

from sklearn.preprocessing import PolynomialFeatures
from sklearn.linear_model import LinearRegression

# Print LB and UB for each column and replace outliers with NaN in the copy
for column in columns_to_process:
    outliers_mask = (df_copy[column] < LB[column]) | (df_copy[column] > UB[column])

    # Replace outliers with NaN in the copy
    df_copy[column] = np.where(outliers_mask, np.nan, df_copy[column])

    # Identify NaN values in the copy
    nan_mask = df_copy[column].isna()

    # Create a copy of the column data to avoid modifying the original array
    poly_data = np.copy(df_copy[column])

    # Replace NaN values with Polynomial regression predictions
    if np.sum(~nan_mask) > 0:
        X = np.arange(len(df_copy[column])).reshape(-1, 1)

        # Extract non-NaN values for training
        X_train = X[~nan_mask].reshape(-1, 1)
        y_train = df_copy[column][~nan_mask]

        # Fit polynomial regression
        poly = PolynomialFeatures(degree=100)
        X_poly = poly.fit_transform(X_train)
        poly_reg = LinearRegression()
        poly_reg.fit(X_poly, y_train)

        # Predict for NaN values
        X_pred = poly.fit_transform(X[nan_mask].reshape(-1, 1))
        poly_data[nan_mask] = poly_reg.predict(X_pred)
        df_copy[column] = poly_data

df_copy_1 = df_copy.copy()

```

```
In [3]: from statsmodels.stats.outliers_influence import variance_inflation_factor

# Assuming df is your DataFrame
columns_to_exclude = ['c1', 'c51', 'c52', 'c53', 'c54']

# Select columns for VIF calculation
columns_for_vif = [col for col in df_copy.columns if col not in columns_to_exclude]

# Create a DataFrame for VIF results
vif_data = pd.DataFrame()
vif_data["Variable"] = columns_for_vif
vif_data["VIF"] = [variance_inflation_factor(df_copy[columns_for_vif].values, i) for i in range(len(columns_for_vif))]

# Display columns with high collinearity (VIF > 20)
high_collinearity_columns = vif_data[vif_data["VIF"] > 20]
print("Columns with High Collinearity:")
print(high_collinearity_columns)
```

```
/usr/local/lib/python3.10/dist-packages/statsmodels/regression/linear_model.py:1781: RuntimeWarning: divide by zero encountered in double_scalars
    return 1 - self.ssr/self.centered_tss
/usr/local/lib/python3.10/dist-packages/statsmodels/regression/linear_model.py:1781: RuntimeWarning: divide by zero encountered in double_scalars
    return 1 - self.ssr/self.centered_tss
/usr/local/lib/python3.10/dist-packages/statsmodels/regression/linear_model.py:1781: RuntimeWarning: divide by zero encountered in double_scalars
    return 1 - self.ssr/self.centered_tss
/usr/local/lib/python3.10/dist-packages/statsmodels/regression/linear_model.py:1781: RuntimeWarning: invalid value encountered in double_scalars
    return 1 - self.ssr/self.centered_tss
/usr/local/lib/python3.10/dist-packages/statsmodels/regression/linear_model.py:1781: RuntimeWarning: divide by zero encountered in double_scalars
    return 1 - self.ssr/self.centered_tss
```

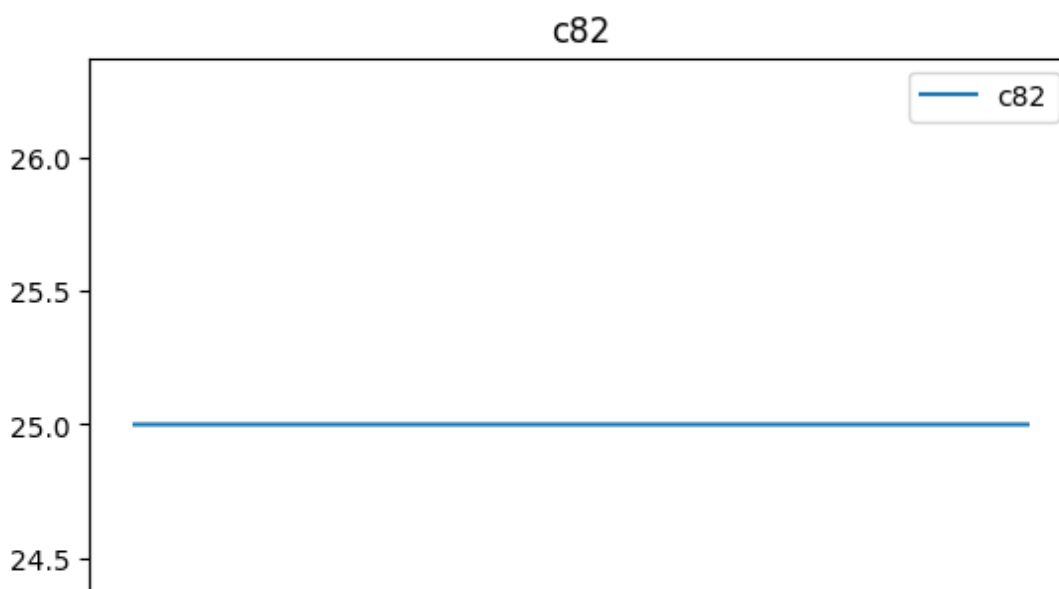
Columns with High Collinearity:

	Variable	VIF
1	c3	1.809614e+02
2	c4	4.254527e+07
3	c5	5.422757e+12
4	c6	1.449501e+12
7	c9	2.082756e+01
..
213	c228	1.192271e+03
215	c235	2.538464e+01
216	c236	9.997501e+01
217	c237	3.128776e+02
220	c241	2.019132e+01

[159 rows x 2 columns]

```
In [4]: problematic_columns = ['c82', 'c88', 'c96', 'c97', 'c102', 'c110', 'c113',  
                               'c133', 'c128', 'c130', 'c207',  
                               'c208', 'c209', 'c210', 'c211', 'c212', 'c213', 'c214',  
                               'c216', 'c217', 'c218', 'c219', 'c220', 'c221']  
  
for column in problematic_columns:  
    print(LB[column],UB[column])  
    plt.figure()  
    plt.plot(df[column], label = column)  
    plt.title(column)  
    plt.legend()  
    plt.show()  
    plt.figure()  
    plt.plot(df_copy[column], label = column)  
    plt.title(column + " after removing all NaN values")  
    plt.legend()  
    plt.show()
```

25.0 25.0



```
In [5]: df_copy.drop(columns = high_collinearity_columns['Variable'].tolist(), inplace=True)
```

```
In [6]: import statsmodels.api as sm

# Assuming df_copy is your preprocessed DataFrame
columns_to_exclude = ['c1', 'c51', 'c52', 'c53', 'c54']
columns_for_mlr_c51 = [col for col in df_copy.columns if col not in columns_to_exclude]

X_c51 = df_copy[columns_for_mlr_c51]
y_c51 = df['c51']

mlr_sm_model_c51 = sm.OLS(y_c51, X_c51).fit()
pred_c51 = mlr_sm_model_c51.predict(X_c51)
print(mlr_sm_model_c51.summary())

columns_for_mlr_c54 = [col for col in df_copy.columns if col not in columns_to_exclude]

X_c54 = df_copy[columns_for_mlr_c54]
y_c54 = df['c54']

mlr_sm_model_c54 = sm.OLS(y_c54, X_c54).fit()
pred_c54 = mlr_sm_model_c54.predict(X_c54)

print(mlr_sm_model_c54.summary())

columns_for_mlr_c52 = [col for col in df_copy.columns if col not in columns_to_exclude]

X_c52 = df_copy[columns_for_mlr_c52]
y_c52 = df['c52']

mlr_sm_model_c52 = sm.OLS(y_c52, X_c52).fit()
pred_c52 = mlr_sm_model_c52.predict(X_c52)

print(mlr_sm_model_c52.summary())

columns_for_mlr_c53 = [col for col in df_copy.columns if col not in columns_to_exclude]

X_c53 = df_copy[columns_for_mlr_c53]
y_c53 = df['c53']

mlr_sm_model_c53 = sm.OLS(y_c53, X_c53).fit()
pred_c53 = mlr_sm_model_c53.predict(X_c53)

print(mlr_sm_model_c53.summary())
```

OLS Regression Results

```

=====
=====
Dep. Variable:          c51    R-squared:
0.673
Model:                  OLS    Adj. R-squared:
0.655
Method:                 Least Squares    F-statistic:
37.00
Date:                  Mon, 13 Nov 2023    Prob (F-statistic):      3.
25e-197
Time:                  14:01:12    Log-Likelihood:
-1909.1
No. Observations:      1025    AIC:
3928.
Df Residuals:          970    BIC:
4200.
Df Model:              54
Covariance Type:       nonrobust

```

```

In [18]: control_param=df_copy_1[['c26', 'c27', 'c28', 'c29', 'c30', 'c31', 'c32', 'c
y51=df_copy['c51']
mlr_model51_con = sm.OLS(y51, control_param).fit()
print(mlr_model51_con.summary())
y52=df_copy['c52']
mlr_model52_con = sm.OLS(y52, control_param).fit()
print(mlr_model52_con.summary())
y53=df_copy['c53']
mlr_model53_con = sm.OLS(y53, control_param).fit()
print(mlr_model53_con.summary())
y54=df_copy['c54']
mlr_model54_con = sm.OLS(y54, control_param).fit()
print(mlr_model54_con.summary())

```

OLS Regression Results

```

=====
=====
Dep. Variable:          c51    R-squared (uncentered):
0.959
Model:                  OLS    Adj. R-squared (uncentered):
0.959
Method:                 Least Squares    F-statistic:
1249.
Date:                  Mon, 13 Nov 2023    Prob (F-statistic):
0.00
Time:                  15:03:56    Log-Likelihood:
-2150.2
No. Observations:      1025    AIC:
4338.
Df Residuals:          1006    BIC:
4432.
Df Model:              19
Covariance Type:       nonrobust

```

```
In [8]: prediction=mlr_model51_con.predict(control_param)

predicted=[]
for i in range(0,prediction.shape[0]):
    if(prediction[i]<5):
        predicted.append('safe')
    elif (prediction[i]<10 and 5<prediction[i]):
        predicted.append('moderate')
    elif (prediction[i]<20 and 10<prediction[i]):
        predicted.append('high')
    else:
        predicted.append('critical')
print(predicted)
```

8/29

[illegible]

```
In [9]: prediction=mlr_model52_con.predict(control_param)

predicted=[]
for i in range(0,prediction.shape[0]):
    if(prediction[i]<5):
        predicted.append('safe')
    elif (prediction[i]<10 and 5<prediction[i]):
        predicted.append('moderate')
    elif (prediction[i]<20 and 10<prediction[i]):
        predicted.append('high')
    else:
        predicted.append('critical')
print(predicted)
```

[illegible]

[illegible]

[illegible]

```
In [10]: prediction=mlr_model53_con.predict(control_param)

predicted=[]
for i in range(0,prediction.shape[0]):
    if(prediction[i]<5):
        predicted.append('safe')
    elif (prediction[i]<10 and 5<prediction[i]):
        predicted.append('moderate')
    elif (prediction[i]<20 and 10<prediction[i]):
        predicted.append('high')
    else:
        predicted.append('critical')
print(predicted)
```

[illegible]


```
ate', 'moderate', 'high', 'high', 'high', 'high', 'high', 'high', 'high', 'high',  
'high', 'high', 'high', 'high', 'high', 'high', 'high', 'high', 'moderate', 'hig  
h', 'high', 'high', 'high', 'high', 'high', 'high', 'high', 'high', 'high', 'hig  
h', 'high', 'high', 'high', 'high', 'high', 'high', 'high', 'high', 'high', 'hig  
h', 'high', 'high', 'high', 'high', 'high', 'high', 'high', 'high', 'high', 'hig  
h', 'high', 'high', 'high', 'high', 'high', 'high', 'high', 'high', 'high', 'hig  
h', 'high', 'moderate', 'high', 'high', 'high', 'high', 'high', 'high', 'high', 'h  
igh', 'high', 'high', 'high', 'high', 'high', 'high', 'high', 'high', 'high', 'hig  
h', 'high', 'high', 'high', 'high', 'high', 'high', 'high', 'high', 'high', 'hig  
h', 'high', 'high', 'high', 'high', 'high', 'high', 'high', 'high', 'high', 'hig  
h', 'high', 'high', 'high', 'high', 'high', 'high', 'high', 'high', 'high', 'hig  
h', 'high', 'moderate', 'high', 'moderate', 'moderate', 'moderate', 'moderat  
ate', 'moderate', 'moderate', 'moderate', 'moderate', 'moderate', 'moderate', 'moderat  
e', 'moderate', 'moderate', 'moderate', 'moderate', 'moderate', 'moderate', 'moderat  
e', 'moderate', 'moderate', 'moderate', 'moderate', 'moderate', 'moderate', 'moderat  
e', 'moderate', 'moderate', 'moderate', 'moderate', 'moderate', 'moderate', 'moderat  
e', 'moderate', 'moderate', 'moderate', 'moderate', 'moderate', 'safe', 'moderate', 'm  
oderate', 'moderate', 'moderate', 'safe', 'safe', 'safe', 'safe', 'safe', 'safe',  
'safe', 'safe', 'safe', 'safe', 'safe', 'safe', 'safe', 'safe', 'safe', 'moderat  
e']
```

```
In [11]: prediction=mlr_model54_con.predict(control_param)

predicted=[]
for i in range(0,prediction.shape[0]):
    if(prediction[i]<5):
        predicted.append('safe')
    elif (prediction[i]<10 and 5<prediction[i]):
        predicted.append('moderate')
    elif (prediction[i]<20 and 10<prediction[i]):
        predicted.append('high')
    else:
        predicted.append('critical')
print(predicted)
```

20/29

[illegible]

```
In [12]: # Extract coefficients and p-values directly from the results
coefficients = mlr_model51_con.params
p_values = mlr_model51_con.pvalues

# Create a DataFrame with variable names, coefficients, and p-values
results = pd.DataFrame({
    'Variable': coefficients.index,
    'Coefficient': coefficients.values,
    'P-value': p_values.values
})

# Sort the DataFrame based on P-values in ascending order
sorted_results = results.sort_values(by='P-value')

# Display the sorted results
print(sorted_results)
```

	Variable	Coefficient	P-value
17	c161	2.075398e-02	6.633554e-26
15	c158	3.115210e-01	2.246108e-17
8	c39	1.300734e+01	1.539583e-12
13	c156	1.302331e-14	2.915443e-09
1	c27	6.476557e-05	2.267978e-04
11	c143	5.596057e-03	1.778786e-03
2	c28	6.952627e-02	9.093442e-03
5	c31	-6.505412e-02	1.565456e-02
4	c30	1.560065e+00	2.283270e-02
6	c32	3.001348e-01	2.601228e-02
14	c157	-9.015735e-02	1.148564e-01
10	c142	7.056150e-02	1.170523e-01
0	c26	-8.898137e-02	1.911171e-01
7	c33	2.200611e-01	4.276409e-01
9	c139	4.072722e-02	4.600851e-01
16	c160	-2.056615e-03	5.504104e-01
3	c29	3.519571e-02	6.021834e-01
12	c155	-5.159549e-03	7.538032e-01
18	c162	-4.314907e-04	8.839341e-01
19	c163	8.646758e-05	9.821910e-01

```
In [13]: # Extract coefficients and p-values directly from the results
coefficients = mlr_model52_con.params
p_values = mlr_model52_con.pvalues

# Create a DataFrame with variable names, coefficients, and p-values
results = pd.DataFrame({
    'Variable': coefficients.index,
    'Coefficient': coefficients.values,
    'P-value': p_values.values
})

# Sort the DataFrame based on P-values in ascending order
sorted_results = results.sort_values(by='P-value')

# Display the sorted results
print(sorted_results)
```

	Variable	Coefficient	P-value
15	c158	3.601033e-01	1.985080e-53
17	c161	1.253529e-02	2.160394e-25
3	c29	-3.679800e-01	2.073346e-18
2	c28	1.407381e-01	1.835697e-17
8	c39	8.674654e+00	1.341318e-14
0	c26	2.989787e-01	1.212993e-12
7	c33	1.079346e+00	2.837637e-10
13	c156	8.378545e-15	4.202238e-10
14	c157	2.072156e-01	3.972648e-09
9	c139	-1.835047e-01	6.302097e-08
1	c27	5.527932e-05	2.811771e-07
5	c31	7.815165e-02	2.205135e-06
6	c32	-3.818035e-01	3.893964e-06
11	c143	5.003893e-03	5.071776e-06
4	c30	1.893196e+00	6.634028e-06
12	c155	-3.974052e-02	8.164644e-05
19	c163	9.290505e-03	9.173242e-05
10	c142	-6.441022e-02	1.926699e-02
16	c160	3.737062e-03	7.587779e-02
18	c162	1.262280e-03	4.845180e-01


```
In [14]: # Extract coefficients and p-values directly from the results
coefficients = mlr_model53_con.params
p_values = mlr_model53_con.pvalues

# Create a DataFrame with variable names, coefficients, and p-values
results = pd.DataFrame({
    'Variable': coefficients.index,
    'Coefficient': coefficients.values,
    'P-value': p_values.values
})

# Sort the DataFrame based on P-values in ascending order
sorted_results = results.sort_values(by='P-value')

# Display the sorted results
print(sorted_results)
```

	Variable	Coefficient	P-value
12	c155	5.178470e-01	1.034787e-95
19	c163	5.646324e-02	1.468468e-25
15	c158	3.630598e-01	2.557733e-13
2	c28	2.412885e-01	3.867750e-11
1	c27	1.503051e-04	3.738285e-10
14	c157	4.201907e-01	7.492538e-08
11	c143	1.238280e-02	3.884922e-07
13	c156	-1.036167e-14	4.652165e-04
8	c39	-7.877056e+00	1.439561e-03
17	c161	6.285676e-03	1.589226e-02
4	c30	1.943844e+00	3.660463e-02
9	c139	-1.293465e-01	8.406650e-02
7	c33	5.991213e-01	1.117304e-01
0	c26	-1.315350e-01	1.545200e-01
3	c29	1.275863e-01	1.640119e-01
16	c160	6.297493e-03	1.780548e-01
18	c162	-4.820621e-03	2.296489e-01
6	c32	1.583405e-01	3.863826e-01
5	c31	2.285048e-03	9.500522e-01
10	c142	1.826447e-04	9.976136e-01

```
In [15]: # Extract coefficients and p-values directly from the results
coefficients = mlr_model54_con.params
p_values = mlr_model54_con.pvalues

# Create a DataFrame with variable names, coefficients, and p-values
results = pd.DataFrame({
    'Variable': coefficients.index,
    'Coefficient': coefficients.values,
    'P-value': p_values.values
})

# Sort the DataFrame based on P-values in ascending order
sorted_results = results.sort_values(by='P-value')

# Display the sorted results
print(sorted_results)
```

	Variable	Coefficient	P-value
12	c155	3.226652e-01	1.353274e-41
15	c158	6.217864e-01	4.050970e-33
19	c163	5.434100e-02	5.371011e-23
17	c161	1.691118e-02	3.021425e-10
2	c28	2.320472e-01	4.672364e-10
13	c156	-1.426820e-14	2.519920e-06
14	c157	3.472687e-01	1.284995e-05
7	c33	1.677560e+00	1.412415e-05
4	c30	3.923866e+00	3.830761e-05
8	c39	-1.017901e+01	5.695392e-05
1	c27	9.421859e-05	1.098802e-04
11	c143	8.362206e-03	7.604396e-04
5	c31	8.128869e-02	2.933691e-02
9	c139	-1.628563e-01	3.329400e-02
10	c142	-1.277989e-01	4.069818e-02
16	c160	8.693567e-03	6.885421e-02
6	c32	-2.080862e-01	2.651805e-01
18	c162	-2.352401e-03	5.659537e-01
3	c29	4.669651e-02	6.178847e-01
0	c26	3.237333e-02	7.314491e-01

```
In [16]: columns_to_exclude = ['c1', 'c51', 'c52', 'c53', 'c54']
columns_for_mlr_c241 = [col for col in df_copy.columns if col not in column

X_c241 = df_copy[columns_for_mlr_c241]
y_c241 = df['c241']

mlr_sm_model_c241 = sm.OLS(y_c241, X_c241).fit()

# Extract coefficients and p-values directly from the results
coefficients241 = mlr_sm_model_c241.params
p_values241 = mlr_sm_model_c241.pvalues

# Create a DataFrame with variable names, coefficients, and p-values
results_c241 = pd.DataFrame({
    'Variable': coefficients241.index,
    'Coefficient': coefficients241.values,
    'P-value': p_values241.values
})

# Filter variables with p-values less than 0.05
significant_results_c241 = results_c241[results_c241['P-value'] < 0.05]

# Display the significant results
print(significant_results_c241)

significant_variable_names = significant_results_c241['Variable'].values
columns_for_mlr_c241_finalb = [col for col in df_copy.columns if col in sig

X_c241 = df_copy[columns_for_mlr_c241_finalb]
mlr_sm_model_c241 = sm.OLS(y_c241, X_c241).fit()
print(len(columns_for_mlr_c241_finalb))
print(mlr_sm_model_c241.summary())
```

	Variable	Coefficient	P-value
8	c20	0.086403	6.272224e-04
10	c22	-0.144736	4.062602e-09
11	c23	0.053989	2.237151e-02
12	c30	-0.603754	1.179912e-02
17	c42	0.358678	4.847498e-03
18	c44	-0.186047	3.084779e-04
21	c60	-0.246794	1.862429e-02
23	c62	-0.238910	3.316080e-02
25	c72	-0.152748	2.925802e-02
33	c137	-0.127683	7.348400e-04
39	c152	16.474173	1.336661e-04
53	c178	0.414866	8.746317e-03
59	c230	0.093507	2.802967e-02
13			

OLS Regression Results

```

=====
=====
Dep. Variable:          c241    R-squared (uncentered):
0.910
Model:                  OLS     Adj. R-squared (uncentered):
0.909
Method:                 Least Squares    F-statistic:
785.6
Date:                   Mon, 13 Nov 2023    Prob (F-statistic):
0.00
Time:                   14:01:13    Log-Likelihood:
-1088.9
No. Observations:      1025    AIC:
2204.
Df Residuals:          1012    BIC:
2268.
Df Model:              13
Covariance Type:       nonrobust
=====
=====

```

	coef	std err	t	P> t	[0.025	0.
975]						

c20	0.0275	0.017	1.595	0.111	-0.006	
0.061						
c22	-0.0956	0.018	-5.277	0.000	-0.131	-
0.060						
c23	0.0297	0.016	1.810	0.071	-0.002	
0.062						
c30	-0.3471	0.202	-1.718	0.086	-0.744	
0.049						
c42	0.2668	0.103	2.589	0.010	0.065	
0.469						
c44	-0.0988	0.037	-2.688	0.007	-0.171	-
0.027						
c60	-0.2225	0.068	-3.259	0.001	-0.357	-
0.089						
c62	-0.2275	0.071	-3.214	0.001	-0.366	-
0.089						
c72	-0.2279	0.055	-4.174	0.000	-0.335	-
0.121						
c137	-0.1155	0.028	-4.055	0.000	-0.171	-
0.060						
c152	9.2060	2.707	3.401	0.001	3.894	1

```

4.518
c178          0.1369      0.042      3.286      0.001      0.055
0.219
c230          0.0120      0.029      0.409      0.682      -0.045
0.069
=====
====
Omnibus:                2586.292   Durbin-Watson:
1.332
Prob(Omnibus):          0.000   Jarque-Bera (JB):      2448394
1.154
Skew:                   25.743   Prob(JB):
0.00
Kurtosis:               758.402   Cond. No.                1.23
e+04
=====
====

```

Notes:

- [1] R^2 is computed without centering (uncentered) since the model does not contain a constant.
- [2] Standard Errors assume that the covariance matrix of the errors is correctly specified.
- [3] The condition number is large, $1.23e+04$. This might indicate that there are strong multicollinearity or other numerical problems.

In [16]: