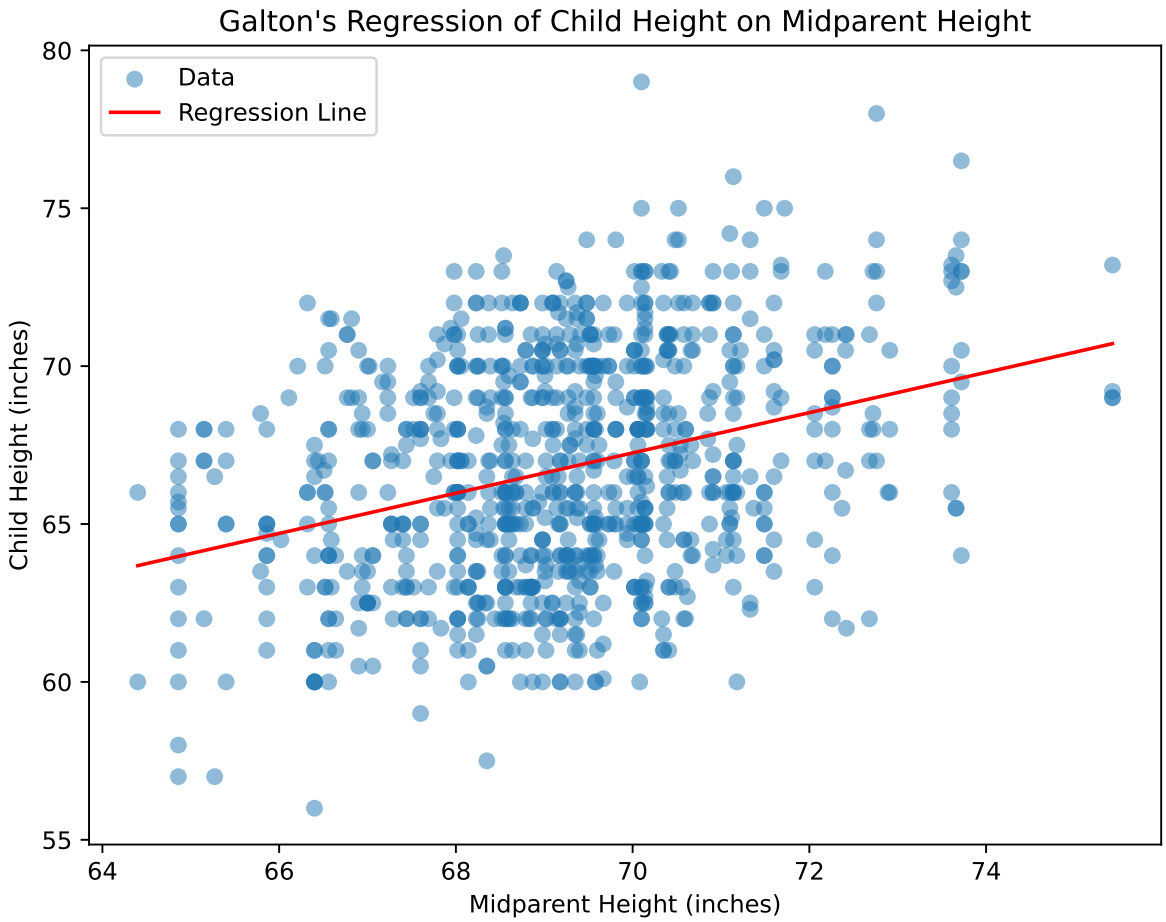
Problem

```
import pandas as pd
import statsmodels.api as sm
import matplotlib.pyplot as plt

# load Galton data
galton = sm.datasets.get_rdataset("GaltonFamilies", "HistData").data

# calculate midparentHeight
galton['midparentHeight'] = (galton['father'] + 1.08 * galton['mother']) / 2

# setup X and Y
X = galton['midparentHeight']
Y = galton['childHeight']

# add constants for X
X = sm.add_constant(X)

# OLS regression
model = sm.OLS(Y, X).fit()
print(model.summary())

# sactter plot
plt.figure(figsize=(8, 6))
plt.scatter(galton['midparentHeight'], galton['childHeight'], alpha=0.5,\
            label='Data')

# regression line
x_vals = pd.Series([min(X['midparentHeight']), max(X['midparentHeight'])])
y_vals = model.params[0] + model.params[1] * x_vals
plt.plot(x_vals, y_vals, color='red', label='Regression Line')

# lable
```

```
plt.xlabel("Midparent Height (inches)")
plt.ylabel("Child Height (inches)")
plt.title("Galton's Regression of Child Height on Midparent Height")
plt.legend()
plt.show()
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:            childHeight   R-squared:                       0.103
Model:                            OLS   Adj. R-squared:                  0.102
Method:                 Least Squares   F-statistic:                     107.0
Date:                Sat, 01 Feb 2025   Prob (F-statistic):           8.05e-24
Time:                        20:07:13   Log-Likelihood:                -2465.0
No. Observations:                 934   AIC:                             4934.
Df Residuals:                     932   BIC:                             4944.
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                     coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const             22.6362      4.265      5.307      0.000      14.266      31.007
midparentHeight    0.6374      0.062     10.345      0.000       0.516       0.758
==============================================================================
Omnibus:                       48.564   Durbin-Watson:                   1.386
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               19.850
Skew:                           0.061   Prob(JB):                     4.89e-05
Kurtosis:                       2.296   Cond. No.                     2.66e+03
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 2.66e+03. This might indicate that there are
strong multicollinearity or other numerical problems.
```

```
/tmp/ipykernel_699/512171542.py:29: FutureWarning: Series.__getitem__ treating keys as positi
  y_vals = model.params[0] + model.params[1] * x_vals
```


Galton's Regression of Child Height on Midparent Height
Data
Regression Line
Child Height (inches)
Midparent Height (inches)
80
75
70
65
60
55
64
66
68
70
72
74