Vasilopoulos Konstantinos

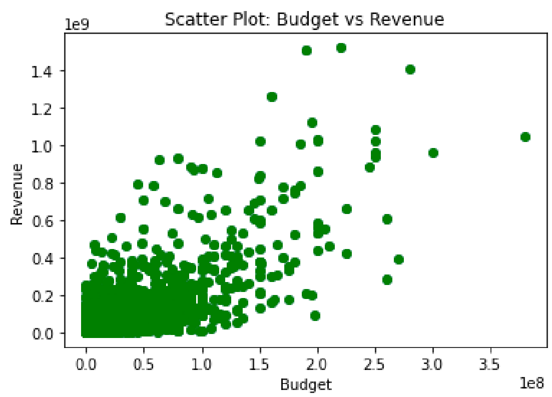
Applied Statistics
Programming Assignment

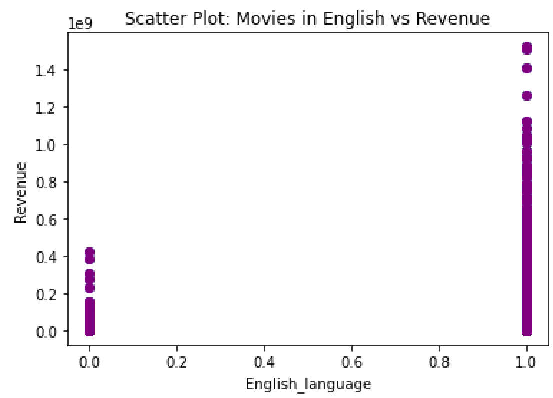
Part 1

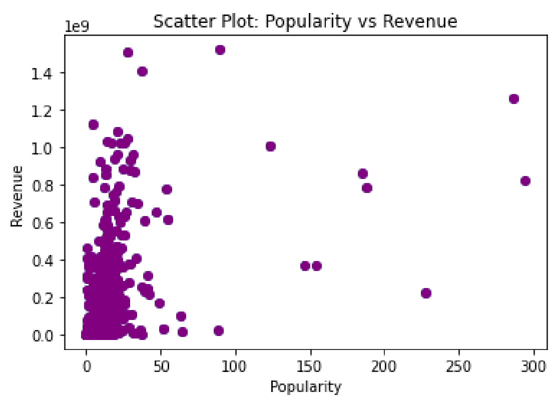
We must calculate the correlation coefficient (r) for every given numerical explanatory variable against the response variable (the one we want to predict = revenue). Then we want to visualize the relationship between them using scatterplots.

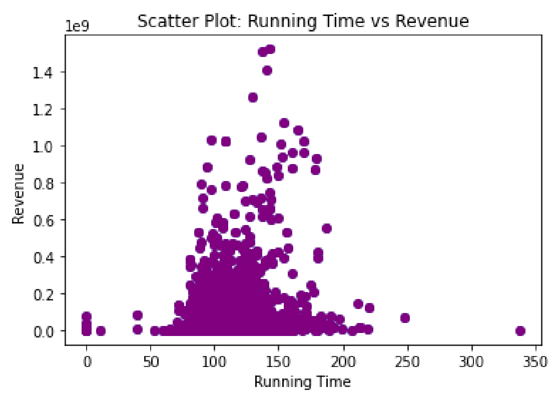
the code (section 1) printed the results below:

Response variable	Explanatory variable	r
Revenue	Budget	0.75296451
Revenue	Popularity	0.46146028
Revenue	English Language	0.14212987
Revenue	Runtime	0.21638013





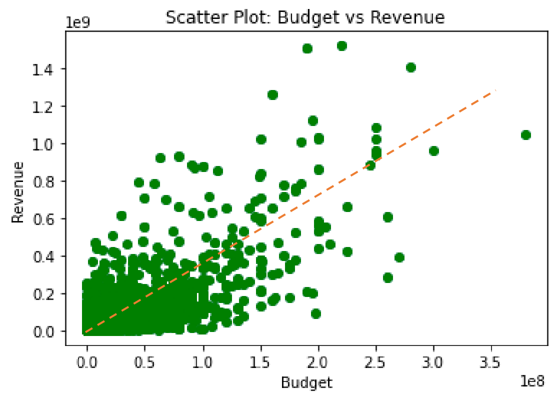




Now that we have the needed results we can decide which one of our explanatory variables (budget, popularity, English language and runtime) is best for predicting a movie's revenue.
In order to choose we take into consideration:

- how strong is the linear relationship between our response and explanatory variable (if there is one)
- which variable has the highest correlation coefficient.

The correct answer is "Budget", which scatterplot is the green one. We can see a strong linear relationship and a high correlation coefficient r = 0.7529645103815287



Part 2

a) The research we are going to conduct has as target the prediction of a movie's revenue according to these five explanatory variables:
   - budget
   - popularity
   - English speaking
   - Runtime
   - Budget larger than the average budget of all movies in dataset (new variable added by us)

   We will use multiple regression with the formula below **(eq 1)**:

$$\boxed{\widehat{revenues} = b_0 + b_1\,\widehat{budget} + b_2\,\widehat{popularity} + b_3\,\widehat{english} + b_4\,\widehat{runtime} + b_5\,avg\_\widehat{budget}}$$

the code (section 2) printed the results below ***(table 1)***:

```
                      Results: Ordinary least squares
=========================================================================================
Model:                     OLS                     Adj. R-squared:            0.622
Dependent Variable:        revenue                 AIC:                       117960.6492
Date:                      2023-06-18 02:33        BIC:                       117996.6834
No. Observations:          2998                    Log-Likelihood:            -58974.
Df Model:                  5                       F-statistic:               987.0
Df Residuals:              2992                    Prob (F-statistic):        0.00
R-squared:                 0.623                   Scale:                     7.1549e+15
-----------------------------------------------------------------------------------------
                     Coef.        Std.Err.       t     P>|t|      [0.025          0.975]
-----------------------------------------------------------------------------------------
Intercept        -27404864.9193 9163611.8143 -2.9906 0.0028 -45372482.4995  -9437247.3391
budget                   2.8212       0.0624 45.2399 0.0000         2.6989         2.9434
popularity         2605826.6099  136410.9221 19.1028 0.0000   2338357.9160   2873295.3037
English_language   1232182.4500 4621108.3055  0.2666 0.7898  -7828688.7995  10293053.6994
runtime             176966.2148   73059.2215  2.4222 0.0155     33714.8223    320217.6072
AVG_budget       -38645159.9161 4873179.1271 -7.9302 0.0000 -48200280.8368 -29090038.9954
-----------------------------------------------------------------------------------------
Omnibus:                     2006.079             Durbin-Watson:                2.028
Prob(Omnibus):               0.000                Jarque-Bera (JB):             62225.134
Skew:                        2.720                Prob(JB):                     0.000
Kurtosis:                    24.646               Condition No.:                268085574
=========================================================================================
* The condition number is large (3e+08). This might indicate            strong
multicollinearity or other numerical problems.
```

Given the results equation 1 can now be written as:

$$\widehat{revenues} = -27404864{,}9193 + 2{,}8212 \cdot \widehat{budget} + 2605826{,}6099 \cdot \widehat{popularity} + 1232182{,}4500 \cdot \widehat{english} + 176966{,}2148 \cdot \widehat{runtime} - 38645159{,}9161 \cdot avg\_\widehat{budget}$$

We are now in position to explain every possible change that our explanatory variables will bring to a movie's revenue. The conclusions we arrive are: (Note that we change one variable at a time and keep the others as constants)

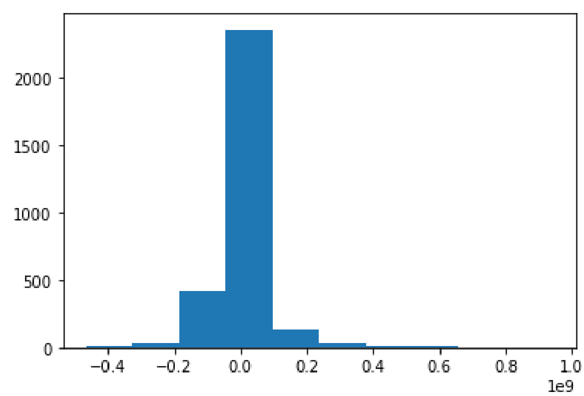
- With every increase at **budget**, **popularity** and **runtime** comes a proportional change in revenues (for instance: with every rise of 1 hour in a movie's runtime we expect its revenues to increase by 176966,2148 dollars.)
- About the **binary variables**, predictions can be made only for movies that have been positive to our demands i.e. that are English and have a larger budget that the average. Although looking at 'AVG_budget' p-value we can see that there are strong evidence against the null Hypothesis, which would be that a movie's revenue does not depend on if its budget is greater than all the movies average of budget (status quo). About the native language we can claim that an English movie can get higher revenues than a 'non-English' one.

- The R^2 for our model is 0.622 and points out that about 62,2% of the data will agree to our predictions. This information tells us that we made a satisfactorily reliable model.

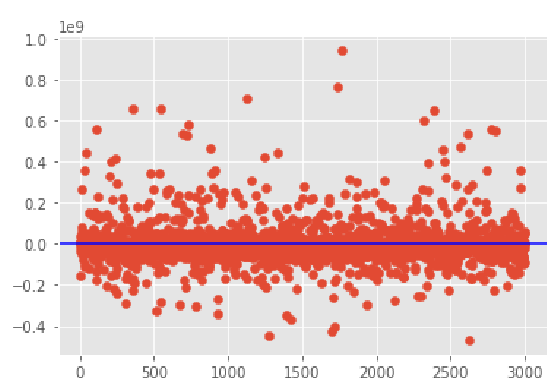
b) We will examine whether our model agree with the conditions for using multiple linear regression. If they are satisfied we can call our predictions accurate and say that the data fits our model.. The conditions are:
   1. Residuals follow nearly normal distribution
   2. Homoscedasticity: same variances within errors
   3. Linear relationship for each of the explanatory (independent) variables with the response variable (dependent) and independent variables need to be unrelated (lack of multicollinearity)

the code (section 3) printed the results below:

1. Given this **residuals histogram** below we can observe that they nearly follow normal distribution



2. Given the **residuals scatterplot** below we can see their variance, which obviously isn't completely consistent among them.



3. Given the **pair plot below** we can see that there is linearity only between revenue (response) and budget (explanatory). As relationship between each independent variable with the dependent one we can easily see that the independent variables are not correlated with each other.

c) We will examine the slope of the variable identified as most predictive ‘budget’. From eq.1 the value of the slope for budget is 2,8212. The meaning of which is that if we keep all the other variables steady and have an increase of the budget by 1 dollar we expect the revenue to increase by 2,8212 dollars

d) In order to answer the question "Which variables are significant for predicting movie revenue?" we must set a confidence level (let's use the most common $\alpha = 0{,}05$) and compare it with the p-value for each explanatory variable. For this operation we are going to use table 1.

Variable	p-value ($P > \|t\|$)	Greater or lower than a	Significant or not
Budget	0,0000	Lower	YES
Popularity	0,0000	Lower	YES
English_language	0,7898	Greater	POSSIBLY NOT (can't say)
Runtime	0,0155	Lower	YES
AVG_budget	0,0000	Lower	YES