Assignment 1

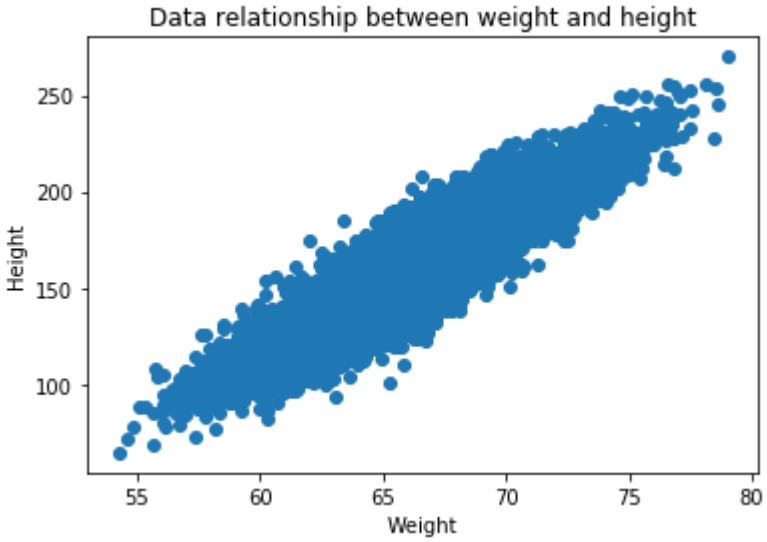
As we saw in my previous report on gender dataset, we have 3 attributes: Gender, Weight and Height, where weight and height are continuous numerical attributes (theoretically they may contain any value within some range, range for weight approximately is from 1 kg to 300 kg and for height is from 0.3 meter to 2.5 meter), and gender is nominal categorical attribute (it is either 0 or 1).

Our task is to predict the height of a person by his weight. To that end we will use two methods:

- In the first method, we ignore the gender of a person and try to make a prediction of height only by weight.
- In the second method, we also consider the gender of a person.

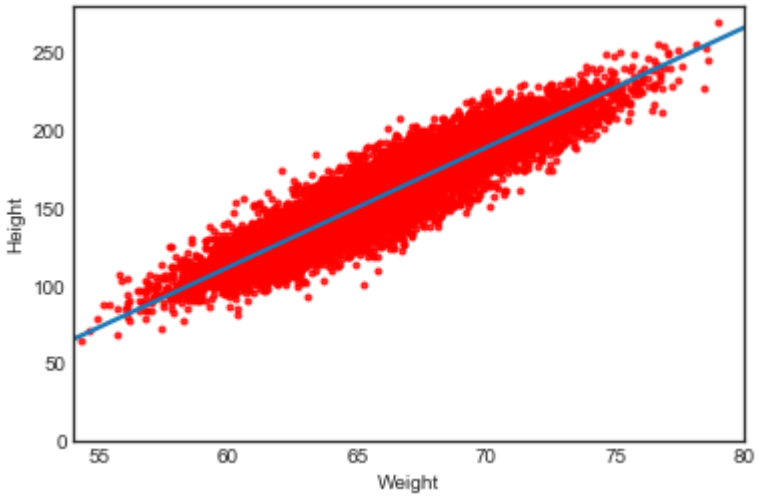
Using weight to predict a person's height:

First, let's look at how weight and height are related to each other.



From above picture, we see that there is linear relationship (linear relationship basically means that when one (or more) independent variables [in our case it is weight] increases (or decreases), the dependent variable increases (or decreases) too) between weight and height, because weight and height values increase in parallel.

Another graphical representation somehow prove linear relationship between Weight and Height.



A little bit about the math.

So we need to construct a model that can be used to estimate Y based on X (In our case Y is Height and X is Weight).

The model can be presented as

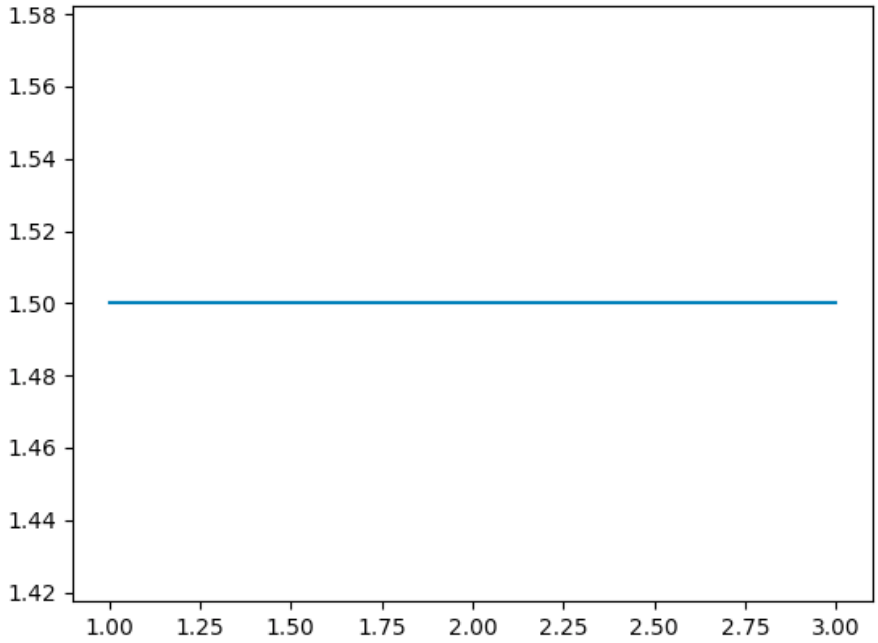
$$h_\theta(x) = \theta_0 + \theta_1 x$$

The theta values are the parameters.

Some quick examples of how we visualize the hypothesis:

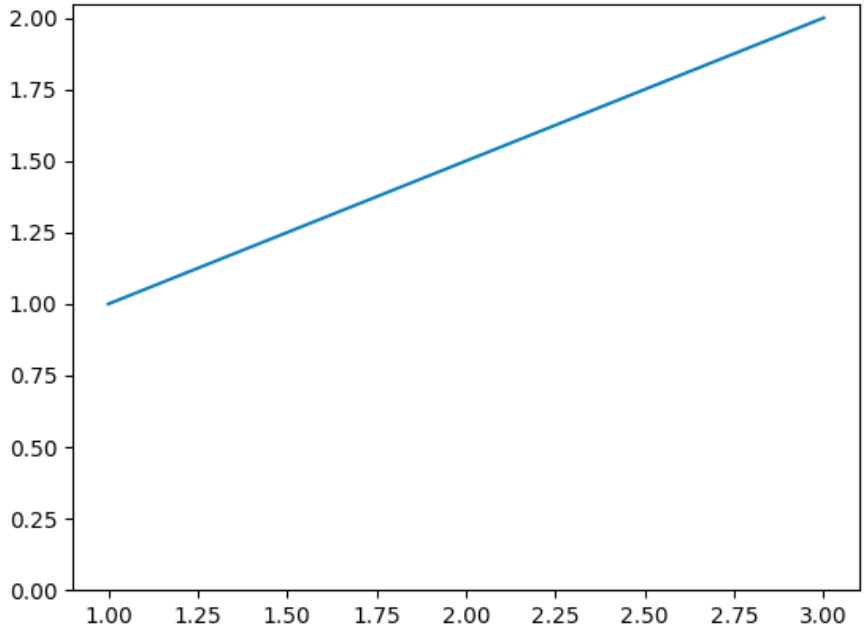
$$\theta_0 = 1.5$$
$$\theta_1 = 0$$

This yields h(x) = 1.5 + 0*x. In this case Y will always be the constant 1.5. This looks like:



How about

$$\theta_0 = 1$$
$$\theta_1 = 0.5$$



The goal of creating a model is to choose parameters, or theta values, so that h(x) is close to y.

Cost Function.

We need a function that will minimize the parameters over our dataset. One common function that is often used is mean squared error, which measure the difference between the estimator (the dataset) and the estimated value (the prediction). It looks like this:

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2.$$

It turns out we can adjust the equation a little to make the calculation down the track a little more simple. We end up with:

$$\frac{1}{2m}\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})^2$$

Gradient Descent.

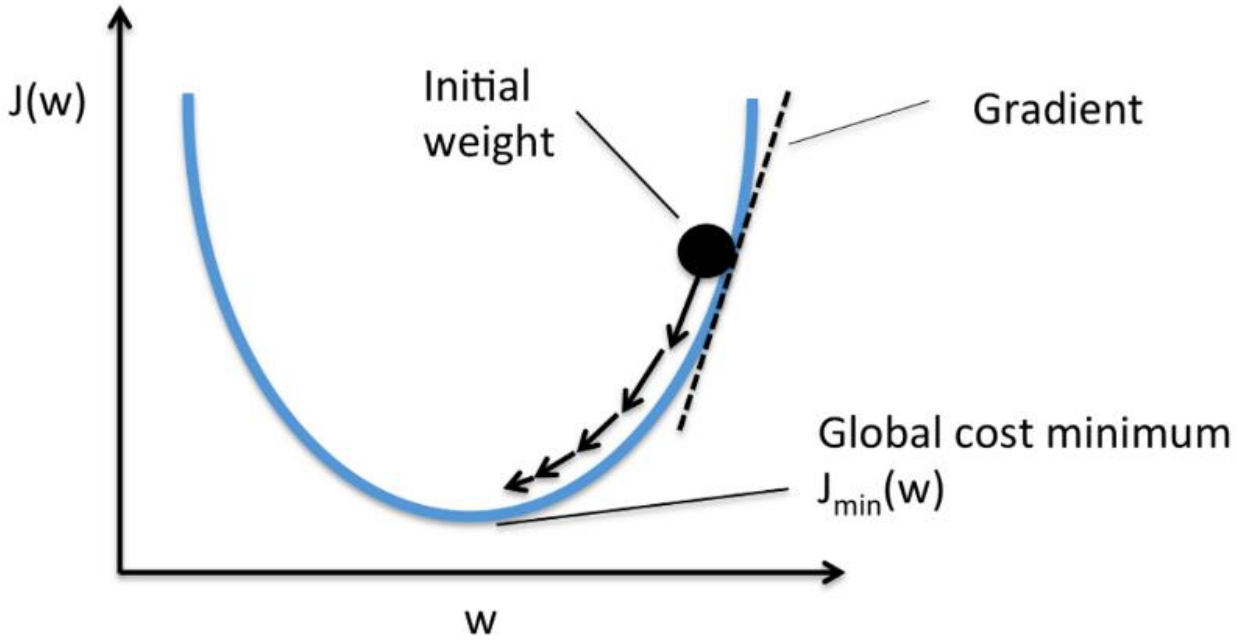
Gradient Descent is a general algorithm for minimizing a function, in this case the Mean Squared Error cost function. Let's denote cost function by $J(\Theta)$.

Gradient Descent basically just change the theta values, or parameters, bit by bit, until we hopefully arrived a minimum.

We start by initializing theta0 and theta1 to any two values, say 0 for both, and go from there. Formally, the algorithm is as follows:

$$\theta_j := \theta_j - \alpha\frac{\partial}{\partial\theta_j}J(\theta_0,\theta_1) \qquad (\text{for } j = 0 \text{ and } j = 1)$$

Where α, alpha, is the learning rate, or how quickly we want to move towards the minimum. If α is too large, however, we can overshoot.



Let's go back to our problem.

So here we have simple linear regression (only one predictor - Weight) and corresponding model gives us the following summary:

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-350.7372	2.111	-166.109	0.000	-354.876	-346.598
Weight	7.7173	0.032	242.975	0.000	7.655	7.780

First we have what's the dependent variable, the model and the method. OLS stands for Ordinary Least Squares and as I mentioned above the method "Least Squares" means that we're trying to fit a regression line that would minimize the square of distance from the regression line.

The coefficient of 7.7173 means that as the Weight variable increases by 1, the predicted value of Height increases by 7.7173. A few other important values are the standard error (is the standard deviation of the sampling distribution of a statistic, most commonly of the mean); the t scores and p-values, for hypothesis test — the Weight has statistically significant p-value; there is a 95% confidence intervals for the Weight (meaning we predict at a 95% percent confidence that the value of Weight is between 7.655 to 7.780).

Other important terms on linear regression models are **R-squared** and **F-statistic.**

Dep. Variable:	Height	**R-squared:**	0.855
Model:	OLS	**Adj. R-squared:**	0.855
Method:	Least Squares	**F-statistic:**	5.904e+04
Date:	Tue, 19 Mar 2019	**Prob (F-statistic):**	0.00
Time:	23:37:00	**Log-Likelihood:**	-39219.
No. Observations:	10000	**AIC:**	7.844e+04
Df Residuals:	9998	**BIC:**	7.846e+04
Df Model:	1		
Covariance Type:	nonrobust		

The hypotheses for the **F-test** of the overall significance are as follows:

Null hypothesis: The fit of the intercept-only model and your model are equal.

Alternative hypothesis: The fit of the intercept-only model is significantly reduced compared to your model.

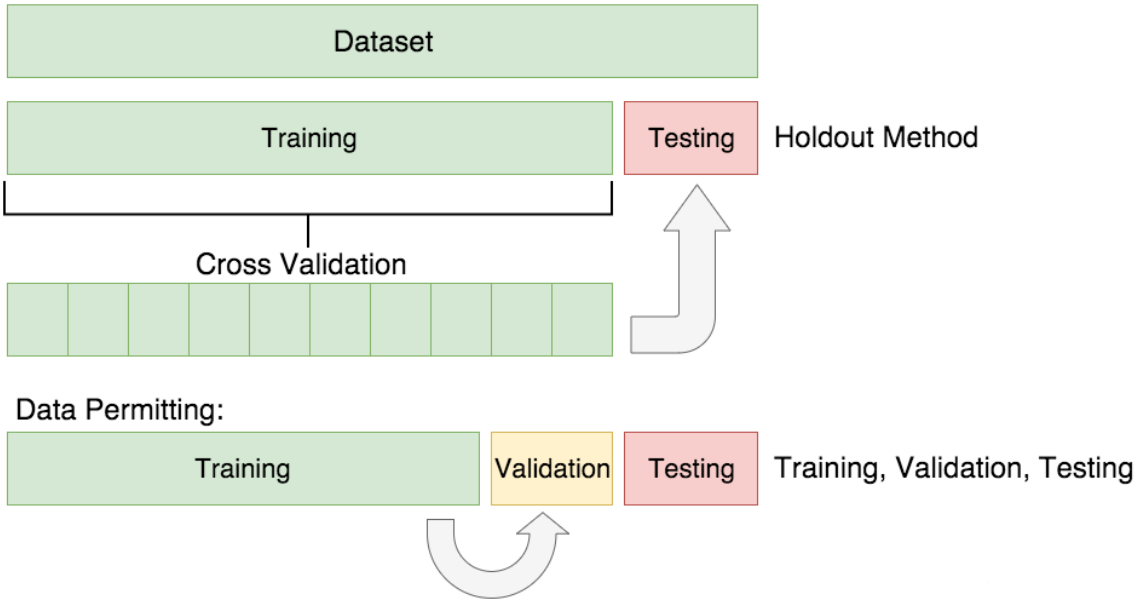
If the P value for the F-test of overall significance test is less than your significance level, you can reject the null-hypothesis and conclude that your model provides a better fit than the intercept-only model. In our case p value << 0.005 so we can reject null hypothesis.

R-squared value — 0.855, meaning that this model explains 85.5% of the variance in our dependent variable.

Cross validation score for this model is equal to 0.5715266817745311

The meaning of cross validation score is the following:

we split our data into k subsets, and train on k-1 one of those subset. What we do is to hold the last subset for test. We're able to do it for each of the subsets.



Using weight and Gender to predict a person's height.

Now let's try fitting a regression model with more than one variable. We will use Weight and Gender features for building such a model.

The summary of that regression model is the follows:

Dep. Variable:	Height	**R-squared:**	0.903
Model:	OLS	**Adj. R-squared:**	0.903
Method:	Least Squares	**F-statistic:**	4.640e+04
Date:	Tue, 19 Mar 2019	**Prob (F-statistic):**	0.00
Time:	23:48:49	**Log-Likelihood:**	-37228.
No. Observations:	10000	**AIC:**	7.446e+04
Df Residuals:	9997	**BIC:**	7.448e+04
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-244.9235	2.299	-106.552	0.000	-249.429	-240.418
Weight	5.9769	0.036	165.973	0.000	5.906	6.048
Gender_Categorical	19.3777	0.277	69.931	0.000	18.835	19.921

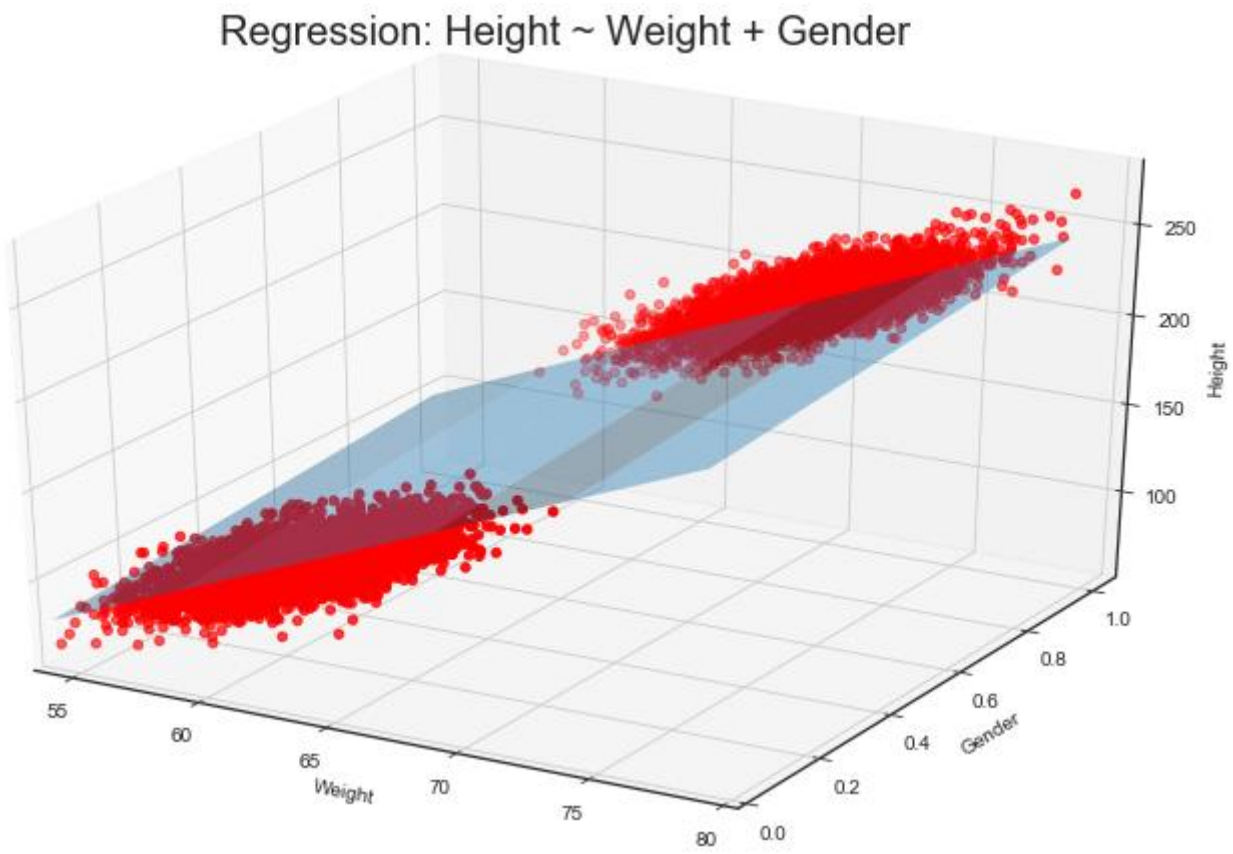
Omnibus:	0.464	**Durbin-Watson:**	2.016
Prob(Omnibus):	0.793	**Jarque-Bera (JB):**	0.447
Skew:	0.016	**Prob(JB):**	0.800
Kurtosis:	3.011	**Cond. No.**	1.53e+03

We can see here that this model has a much higher R-squared value — 0.903, meaning that this model explains 90.3% of the variance in our dependent variable. Whenever we add variables to a regression model, R^2 will be higher, but this is a pretty high R^2. We can see that both Weigh and Gender are statistically significant in predicting the person height value; not surprisingly.

Also the coefficients are changed, Intercept now is equal to -244.9235, weight is 5.9769 and we have new coefficient, corresponds to Gender feature which is 19.3777. All our predictors (Weight and Gender) are statistically significant which mean that both weight and gender are important in regression model.

Cross validation score for this model is equal to 0.7328066585863953

And the graph of the predictor model looks as follows:



Using Interaction term to predict a person's height.

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-246.0133	3.350	-73.443	0.000	-252.579	-239.447
Weight	5.9940	0.053	114.103	0.000	5.891	6.097
Gender_Categorical	21.5144	4.785	4.496	0.000	12.134	30.895
Weight:Gender_Categorical	-0.0323	0.072	-0.447	0.655	-0.174	0.109

As we can see there is no synergy between predictors because p-value for Weight*gender is equal to 0.655, which means that this predictor is not statistically significant.

Conclusion.

Taking into account statistical significance of the features, R-squared values and cross validation scores I intend to choose simple multiple regression model as the final model for predicting person's height. So the final model is the following:

$$\text{Height} = -244.9235 + 5.9769 * \text{Weight} + 19.3777 * \text{Gender}$$