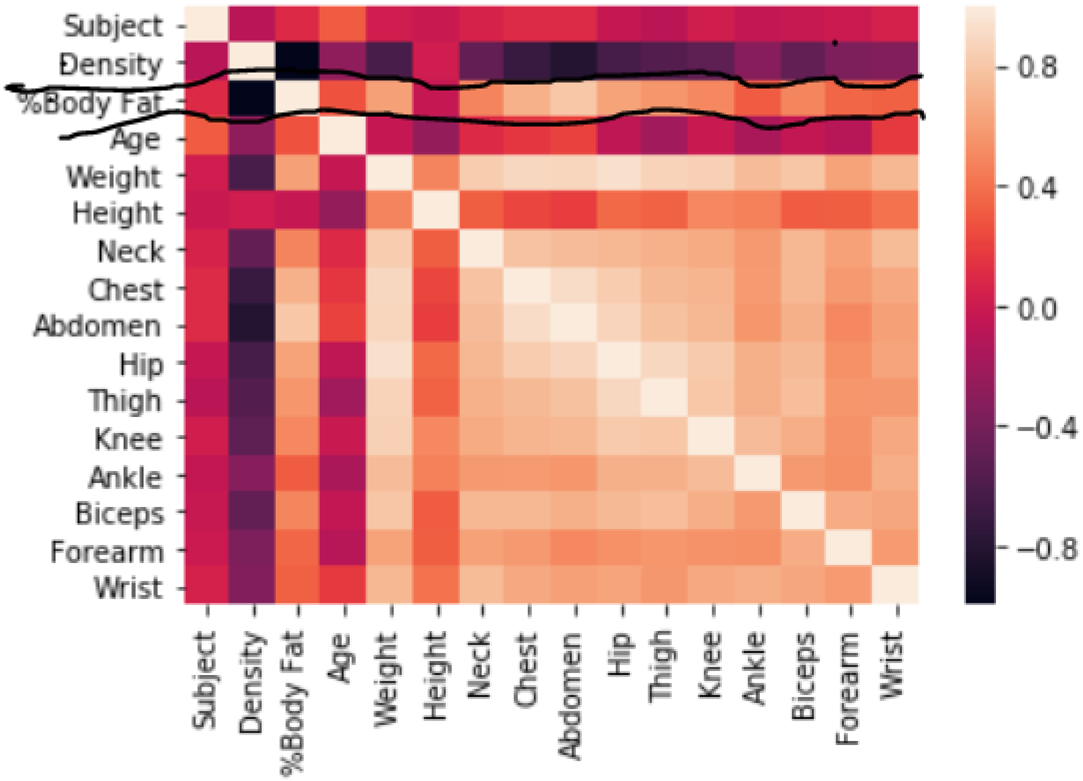
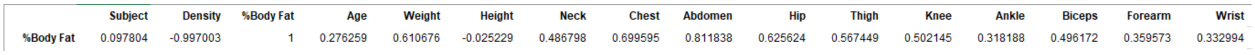
Yusuf Ahmed
10130098
January 23rd 2019
CISC 351 – Assignment 1

1) The first step is to identify which variable to add to the model, which will be the one with the highest linear correlation coefficient with body fat percentage.





	Subject	Density	%Body Fat	Age	Weight	Height	Neck	Chest	Abdomen	Hip	Thigh	Knee	Ankle	Biceps	Forearm	Wrist
%Body Fat	0.097804	-0.997003	1	0.276259	0.610676	-0.025229	0.486798	0.699595	0.811838	0.625624	0.567449	0.502145	0.318188	0.496172	0.359573	0.332994

This plot displays correlation through the colour scheme on the right. The lighter shades indicate higher correlation. The table beneath it highlights the number for each correlation, in very small text. The plot shows that body fat percentage is most highly correlated with:

1. Density
2. Abdomen
3. Chest
4. Hip

The first model will use "Density" as the only variable as it has the highest correlation. The second model will use "Abdomen" as well as "Density" as a two-regressor model. Both models are linear regression models. "Abdomen" is the variable with the highest linear correlation coefficient following "Density". The fitness of the model can be determined by the R^2 and adjusted R^2 values.

The models are linear and follow the format:

$$Y = B_1X + B_2X + B_0$$

Model 1: $Body\ Fat = -444.1 * (Density) + 487.8$

Model 2: $Body\ Fat = -438.3 * (Density) + 0.013\ (Abdomen) + 480.6$

Model	R^2	Adjusted R^2
Model 1	0.898	0.994
Model 2	0.948	0.994

There appears to be no change to the R^2 value with the addition of "Abdomen" to the model. There are other ways to evaluate the two models:

Model	Log-Likelihood	AIC	BIC	Omnibus
Model 1	-245.51	495.0	502.1	409.88
Model 2	-243.63	493.3	503.8	406.73

These parameters also don't indicate a large impact resulting from adding "Abdomen" to the model. The AIC and Log-likelihood indicate Model 2 is slightly stronger, whereas the BIC and Omnibus values point towards Model 1 as the stronger model. It should be noted that in Model 2, Abdomen has a 95% confidence interval range of [0:0.025]. the presence of 0 in the 95% confidence interval indicates that "Abdomen" can be dropped from the model.

2) A model has been built to predict body fat percentage using *Abdomen, Chest, Thigh, Hip* and *Knee* as variables. The resulting model looks like:

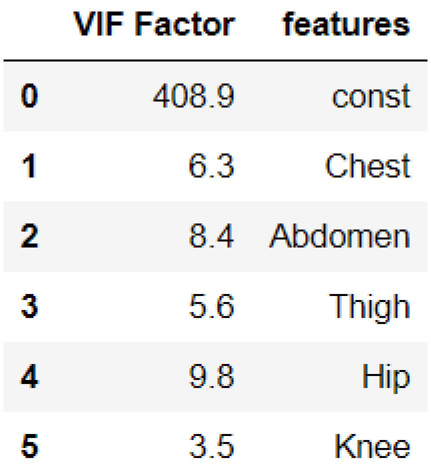
$$\%Body\ Fat = -8.78 - 0.19(Chest) + Abdomen + 0.22(Thigh) - 0.46(Hip) - 0.35(Knee)$$

Each variable will be evaluated to see if it is significant at the 95% level. This can be done by looking at the confidence interval for each variable. Alternatively, the p-value will provide the same information. We are looking for p-values lower than 0.05 (95% confidence interval).

Variable	95% CI Low	95% CI High	P-value
Y-intercept	-20.40	2.84	0.14
Chest	-0.36	-0.02	0.03
Abdomen	0.85	1.16	0
Thigh	-0.04	0.48	0.09
Hip	-0.71	-0.21	0
Knee	-0.79	0.10	0.13

The results show that the y-intercept, *thigh* and *knee* are not statistically significant at the 95% level.

Yusuf Ahmed
10130098
January 23rd 2019
CISC 351 – Assignment 1



	VIF Factor	features
0	408.9	const
1	6.3	Chest
2	8.4	Abdomen
3	5.6	Thigh
4	9.8	Hip
5	3.5	Knee

Looking at the VIF score for each feature, it appears that multicollinearity exists. There is no correlation for *knee* and *thigh. Chest, abdomen* and *hip* need to be investigated further. The y-intercept, *const* is not useful due to it's large VIF factor.