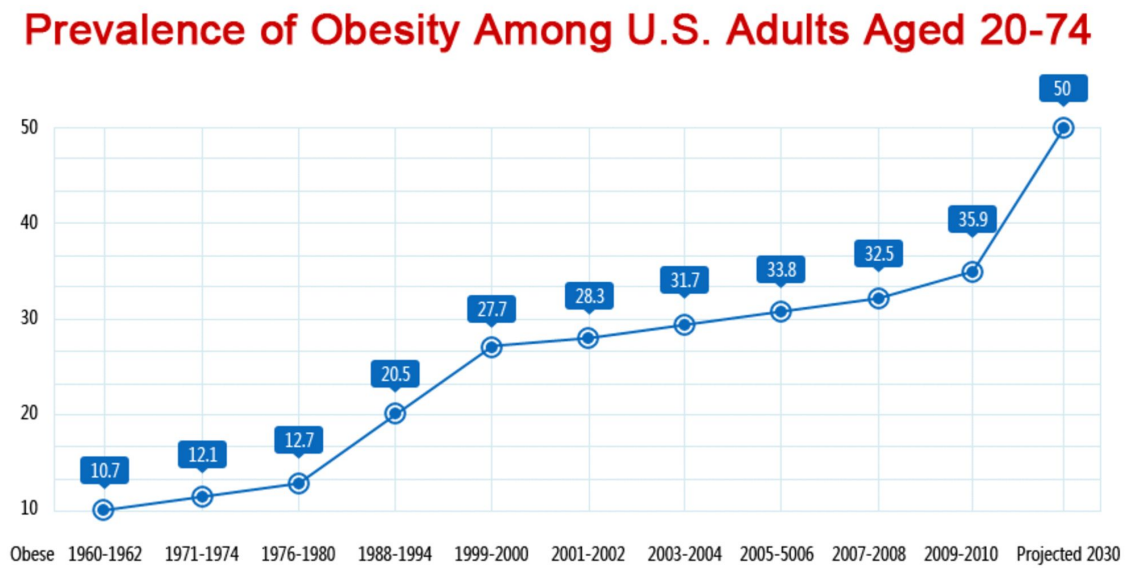
Predictions of Body Types
Or
“Can anyone trust a dating profile?”

Steve Lamont
Codecademy Machine Learning

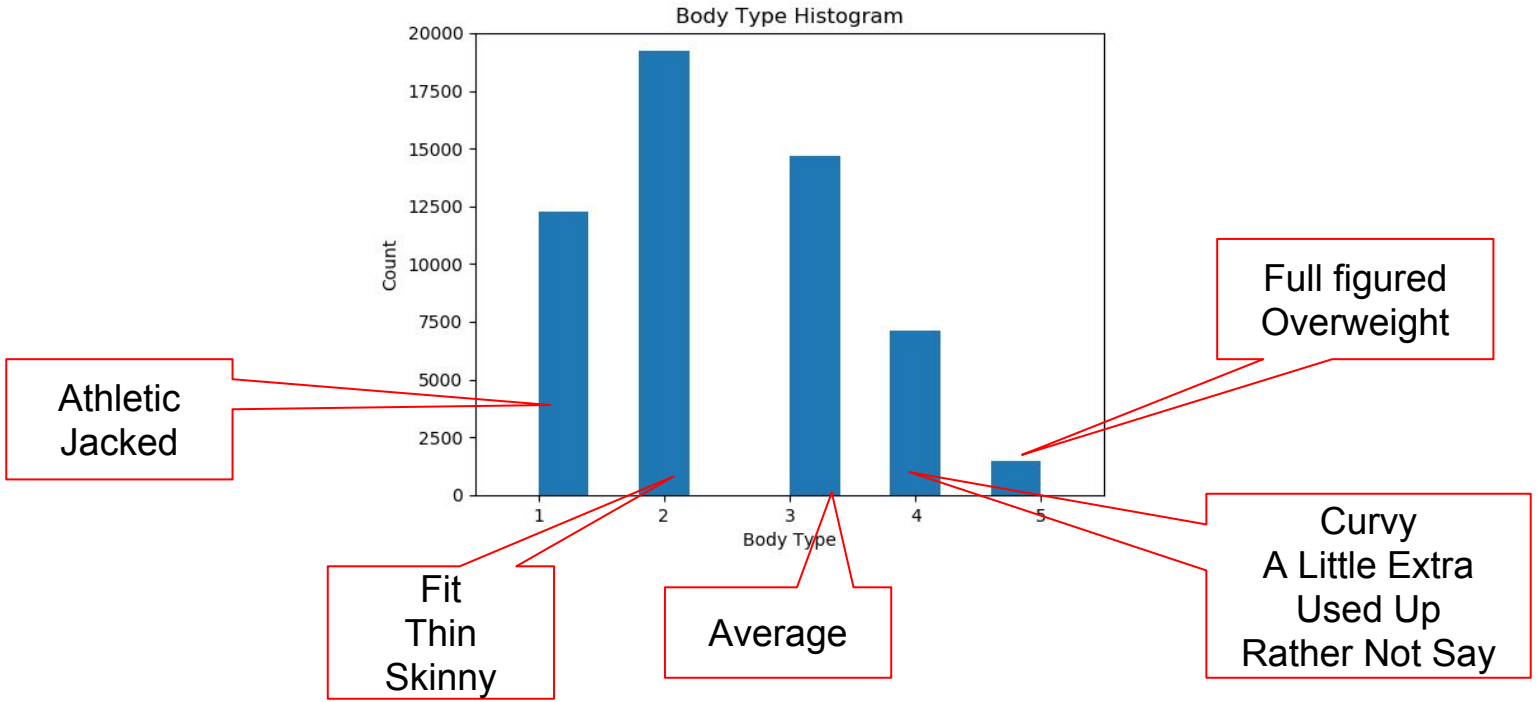
Most of us are aware of the growing obesity problem



Derived from NHANES data (http://www.cdc.gov/nchs/data/hestat/obesity_adult_09_10/obesity_adult_09_10.html#table1)

Our question: Can we apply life factors to predict obesity levels?

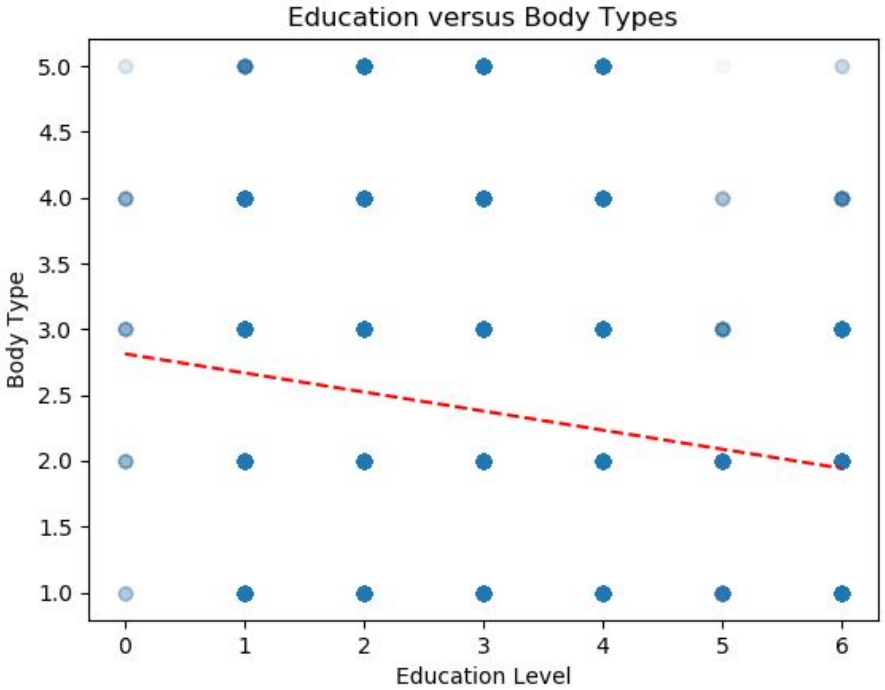
The OK Cupid data set provides a wide set of responses for different body types, which we have grouped into 5 categories



Created extra columns to translate choices into numeric scores for analysis

Example: Body Type	
Description	**Assigned Numeric Score**
Athletic	1
Jacked	1
Fit	1
Thin	2
Skinny	2
Average	3
Curvy	4
A Little Extra	4
Used Up	4
Rather Not Say	4
Full Figured	5
Overweight	5

First let's try to look at education as a predictor of body type



Hypothesis is that higher education levels would lead to better awareness of harm of being overweight and understanding of healthy choices, thus less chances of being overweight

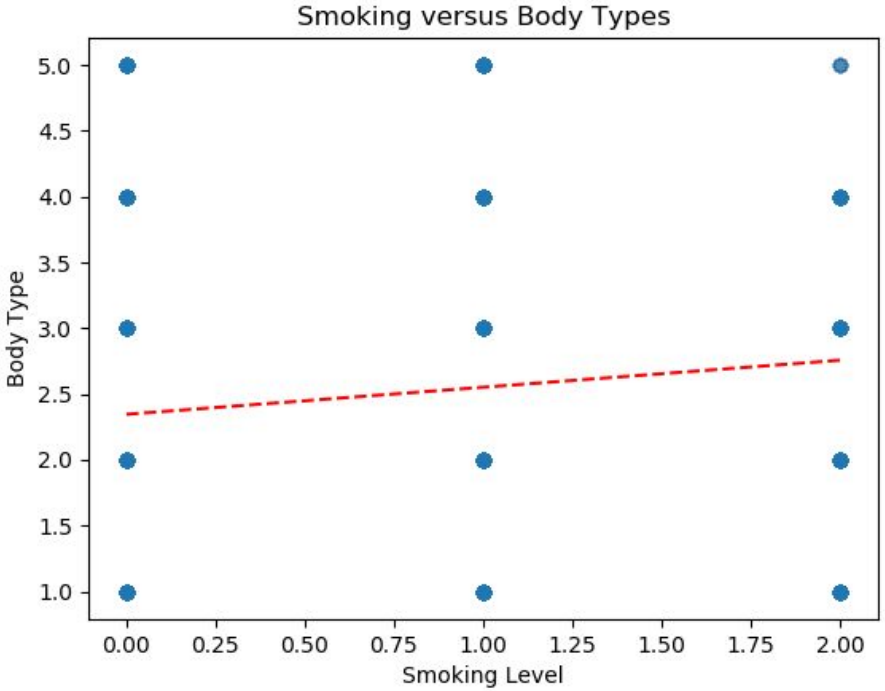
Education levels range from less than High School (0) to PhD (6)

Body type ranges from athletic (1) to overweight (5)

There is a slight inverse correlation, as we might expect. More education means lighter body types.

However, R^2 is only 0.02, so we cannot draw a strong conclusion from this.

Next let's check for smoking versus body type



Hypothesis is that smokers have unhealthy lifestyles and are more likely to be overweight

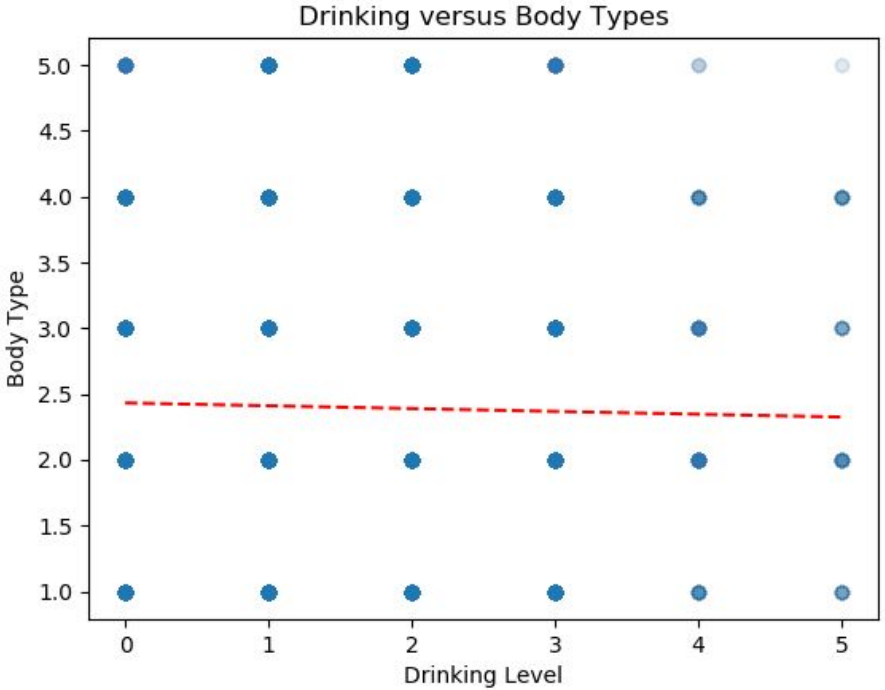
Smoking levels range from None (0) to regular smoker (2)

Body type ranges from athletic (1) to overweight (5)

There is a slight correlation of smoking to heavier body types

However, R^2 is only 0.01

Maybe drinking level will help predict weight?



Hypothesis is that heavier drinkers have unhealthy lifestyles and are more likely to be overweight

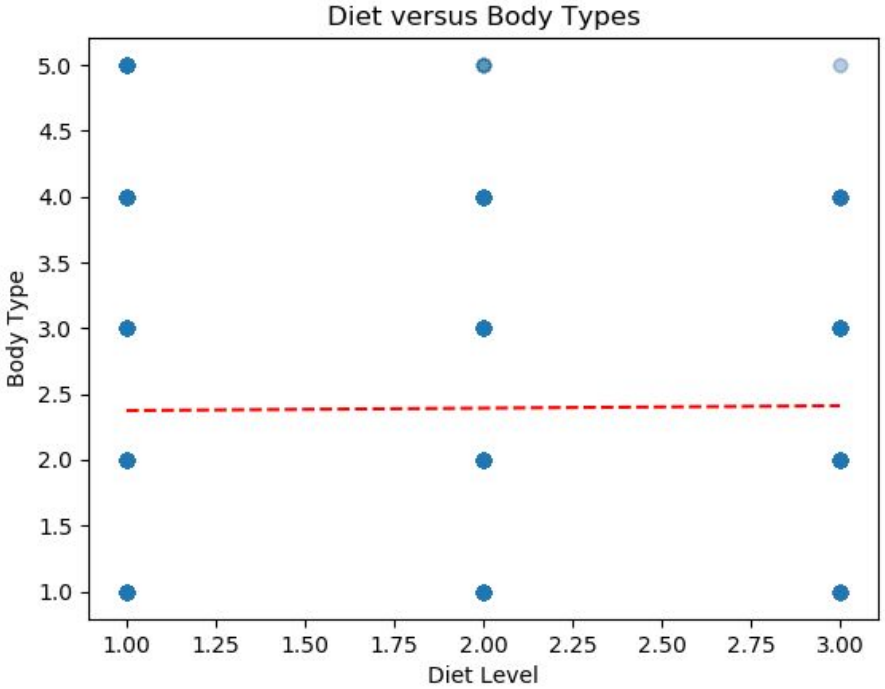
Drinking levels range from None (0) to desperately (5)

Body type ranges from athletic (1) to overweight (5)

There appears to be a slight inverse correlation

However, R^2 is negligible

Perhaps the diet type will affect body type?



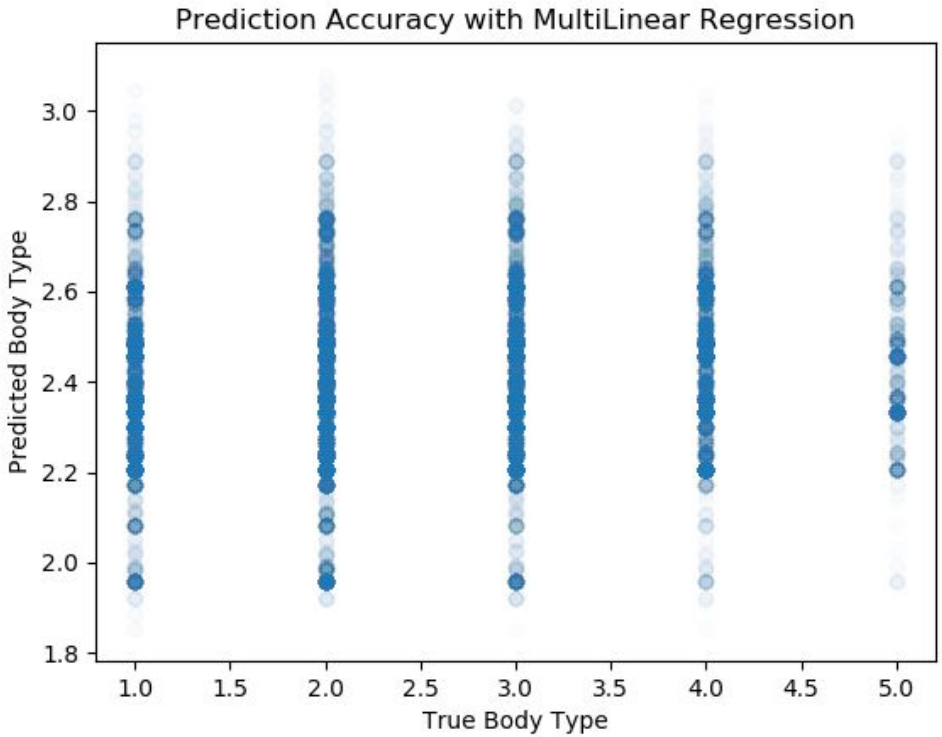
Hypothesis is that people with better diets (vegetarian, vegan) are less likely to be overweight

Diet levels range from Eat Everything (1) to vegetarian/vegan (3)

Body type ranges from athletic (1) to overweight (5)

There is no correlation. The R^2 is close to 0

Even with multilinear regression, our prediction of body type did not improve by much



Multilinear Regression used education, diet, drinking level, and smoking to try to predict body type

Graph shows predicted body type versus true body type

Correlation is still very low, with R^2 of 0.022. Not much of an improvement over simple linear regression based on education

Interesting that smoking showed the highest coefficient (0.15), followed closely by education (-0.13) -- in contrast to education having the higher R^2 in simple linear regression.

What can we take away from this?

1. Most people likely "**fibbed**" about their body type. The distribution on Slide 2 was a giveaway, that if about 40% of Americans are obese today, too few marked that category on their dating profile. This is a dating site, after all, and most people expect to see some stretching of the truth.
2. Assignment of indices for different body descriptions is somewhat **arbitrary**. Changes in the classifications for any of the categories *might* have improved the correlations.
3. A study of correlations such as this would benefit from having a **range of variables** rather than working from a discrete set of choices. For example we would have more gradients if we had the BMI (body mass index) for each participant, and if we had the calorie, alcohol volume, or number of cigarettes intake per day for each participant. The relationships are more likely to be linear than discrete.
4. People's body types may be **individual to their personality, history, and genetics**, and thus we may not be able to find any reliable external predictors.
5. Multilinear Regression is **unlikely to magically improve correlations** when the correlation for each of the underlying Linear Regression is already weak. It is hard to make meaningful data out of garbage. The benefit of sticking with Linear Regression is it is easier to describe and visualize, and allows for testing hypotheses one step at a time.