OkCupid Capstone: Date-A-Scientist

M. Vicky Moya

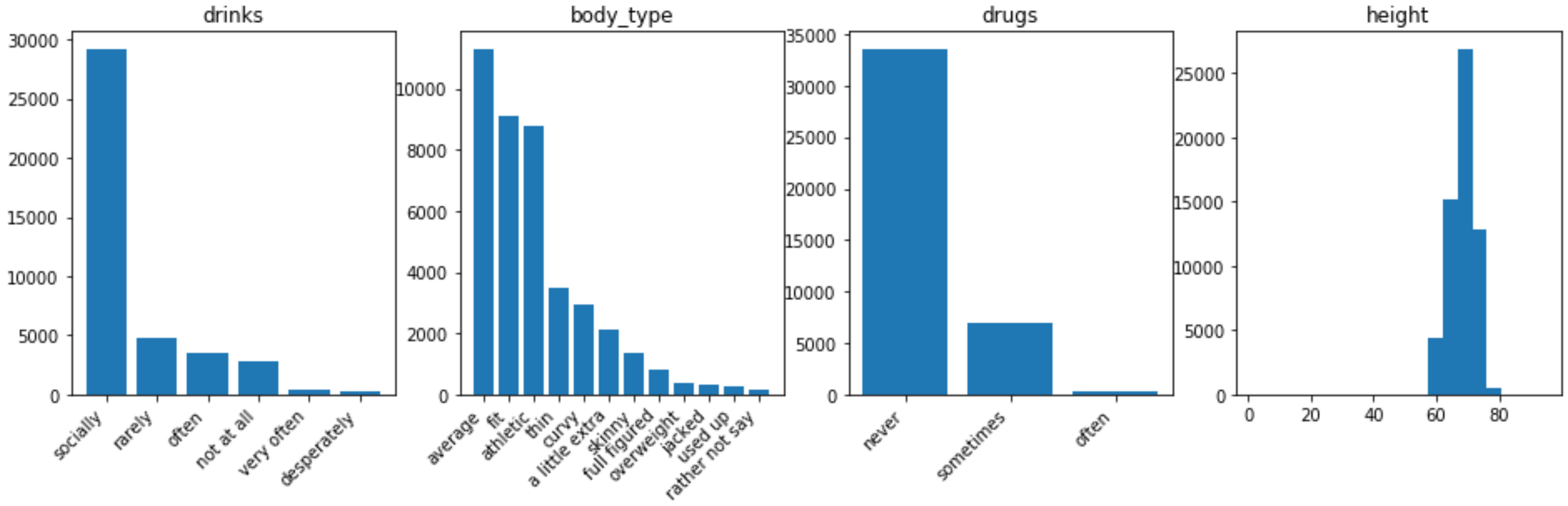
A little bit about the data

The OkCupid dataset is very interesting and has many useful elements, but the have some inherent issues that were fascinating to consider while trying to formulate questions around the data.

First, we'll take a look at some of the data to get a sense of what we're working with.

Summaries of a few categories:

The drinks, body_type, drugs, and height columns are plotted here



Summaries of a few categories:

Most of the features in the dataset are categorical.

Several of the features appear to have pretty strong biases for certain categories. This will make classification difficult as there a fewer examples for certain categories.

Summaries of a few categories:

Given how much data there is on health and lifestyle, I was interested to see if certain lifestyle features correlate with the self-reported health status. In other words, are certain types of lifestyles good at predicting what someone's physique is?

As a proxy for physique, I chose to look at body_type as the target.

Does lifestyle predict body type?

The initial lifestyle features I chose to include for this classification model were:

* diet
* drinks
* drugs
* smokes

codified

* height
* how_active (generated by searching essays for keywords)

Generating the “how_active” column

After all essays were combined, fitness keywords were searched for in the essays. Each appearance of a keyword was counted to generate a semi-quantitative assessment of activity level.

```
sports = ['run', 'running', 'cycling', 'cycle', 'bike', 'swimming', 'swim',
          'climbing', 'climb', 'rowing', 'row', 'surfing', 'surf', 'work-out',
          'gym', 'workout', 'exercise', 'exercising', 'work out', 'working out', 'active',
          'keep active', 'play sports', 'playing sports', 'hockey', 'soccer', 'football',
          'basketball', 'tennis', 'rugby', 'physical activity', 'keep active', 'keeping active', 'kayaking',
          'kayak', 'hiking', 'hike', 'sporty', 'athletic', 'yoga', 'lifting weights', 'weight training',
          'cross-fit', 'crossfit', 'staying fit', 'keep fit', 'stay fit', 'keep fit']
```

Correlations in the data

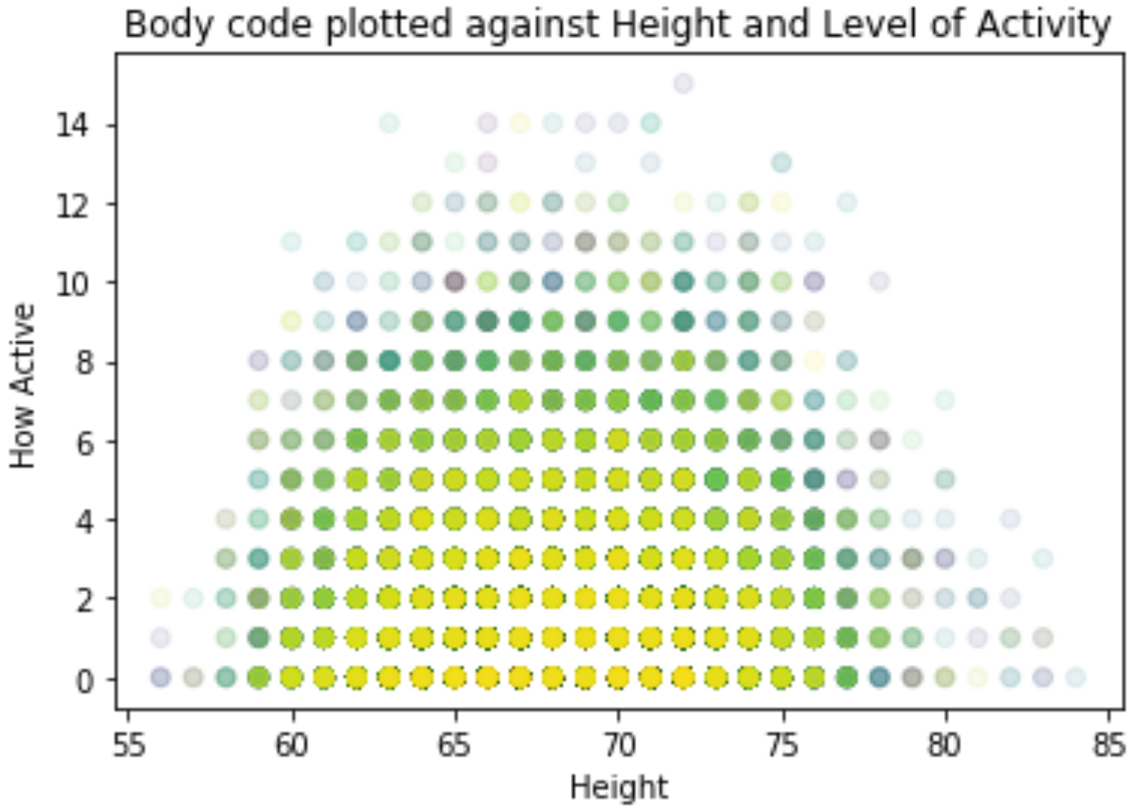
I first looked for inherent correlations of these features:

	height	body_code	diet_code	drinks_code	smokes_code	drugs_code	how_active
height	1.000000	0.176685	0.016883	0.032851	0.044325	0.073398	0.025637
body_code	0.176685	1.000000	0.002569	0.011691	-0.100698	-0.058719	0.203696
diet_code	0.016883	0.002569	1.000000	0.083773	-0.014632	-0.054870	-0.036203
drinks_code	0.032851	0.011691	0.083773	1.000000	0.172119	0.217664	-0.023986
smokes_code	0.044325	-0.100698	-0.014632	0.172119	1.000000	0.363222	-0.135614
drugs_code	0.073398	-0.058719	-0.054870	0.217664	0.363222	1.000000	-0.048569
how_active	0.025637	0.203696	-0.036203	-0.023986	-0.135614	-0.048569	1.000000

height, smokes, and how_active appear to correlate the best

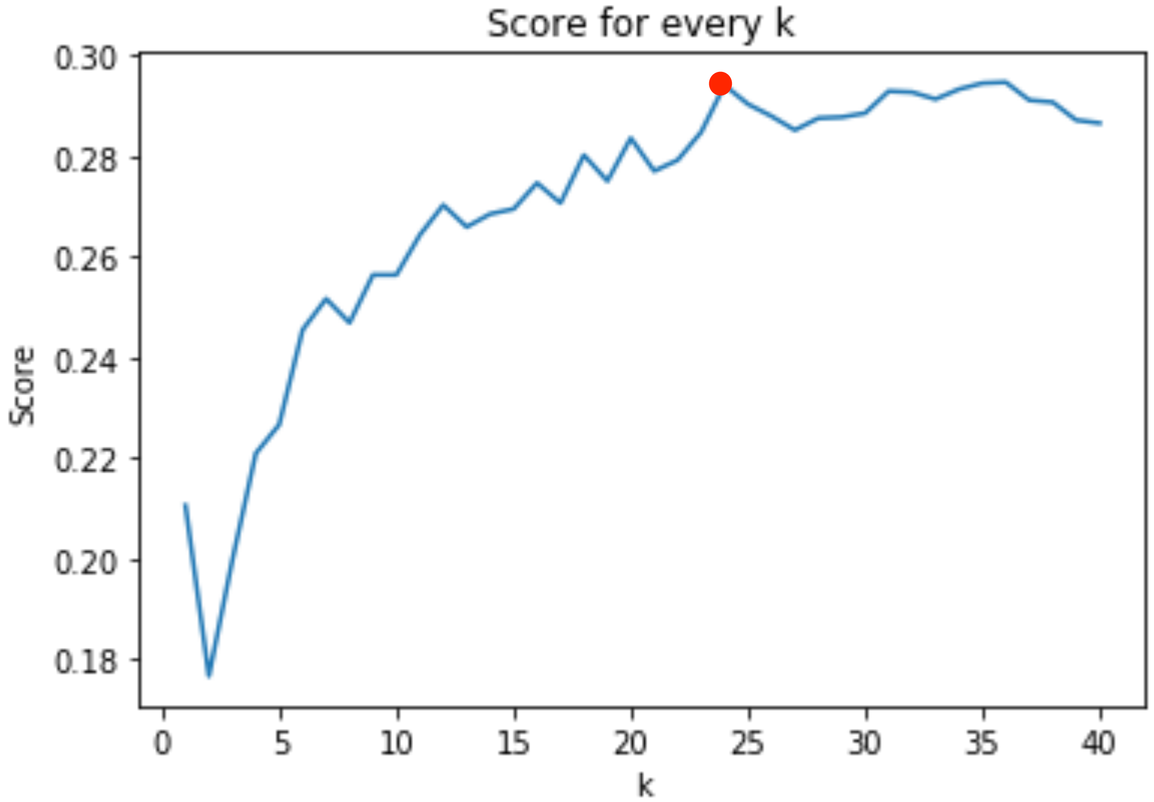
Correlations in the data

To take a quick look at these features against body_type, they were plotted:



Predictive power of height, smokes, and how_active

I first generated a K Nearest Neighbors classifier to predict body_type.
To determine the best k to use for model, I tested multiple values:



Highest accuracy (almost 0.30) was achieved with a k of 35, for this training dataset, but I went with 25, to minimize neighbor noise

Predictive power of height, smokes, and how_active

Even with the most optimal k, the classifier did not perform much better than just guessing by chance.

```
Training data accuracy: 0.3277557100297915; Test data accuracy: 0.2903674280039722
Training data precision: 0.4010061527385855; Test data precision: 0.11127158183089876
Training data recall: 0.1462470853247271; Test data recall: 0.1091577890654752

F1 Scores for training data:
rather not say: 0.085714, used up: 0.100000, overweight: 0.054054, full figured: 0.036176, a little extra: 0.099230,
curvy: 0.218917, average: 0.422810, thin: 0.049277, skinny: 0.042105, fit: 0.299207, athletic: 0.357569, jacked: 0.09
5238

F1 Scores for test data:
rather not say: 0.000000, used up: 0.000000, overweight: 0.000000, full figured: 0.017544, a little extra: 0.006329,
curvy: 0.166667, average: 0.397059, thin: 0.013100, skinny: 0.000000, fit: 0.239922, athletic: 0.332696, jacked: 0.00
0000

Proportions of each class in the test data:
rather not say: 0.002979, used up: 0.008342, overweight: 0.009335, full figured: 0.020060, a little extra: 0.057001,
curvy: 0.072890, average: 0.266336, thin: 0.085998, skinny: 0.033565, fit: 0.222840, athletic: 0.213505, jacked: 0.00
7150
```

Predictive power of height, smokes, and how_active

So it looks like the model may have overfit the training data.

Interestingly, the model seems (slightly) better than chance at predicting when someone has a “fit” or “athletic” body type. This is probably where the “how_active” feature comes through the most.

It also does better than chance at correctly classifying “average” body types. This is likely due to the larger number of “average” classifications in the dataset.

Predictive power of height, smokes, and how_active

Determining the optimal k manually before generating the classifier model takes a little time, so I wanted to try a more automated way of generating an optimized classifier through an SVM:

```
Training data accuracy: 0.31330685203574976; Test data accuracy: 0.30566037735849055
Training data precision: 0.07892696047175195; Test data precision: 0.07665873789134854
Training data recall: 0.10532348161015544; Test data recall: 0.10398016095920043

F1 Scores for training data:
rather not say: 0.000000, used up: 0.000000, overweight: 0.000000, full figured: 0.000000, a little extra: 0.000000,
curvy: 0.000000, average: 0.428015, thin: 0.000000, skinny: 0.000000, fit: 0.186933, athletic: 0.371407, jacked: 0.00
0000

F1 Scores for test data:
rather not say: 0.000000, used up: 0.000000, overweight: 0.000000, full figured: 0.000000, a little extra: 0.000000,
curvy: 0.000000, average: 0.418303, thin: 0.000000, skinny: 0.000000, fit: 0.164349, athletic: 0.373198, jacked: 0.00
0000

Proportions of each class in the test data:
rather not say: 0.002979, used up: 0.008342, overweight: 0.009335, full figured: 0.020060, a little extra: 0.057001,
curvy: 0.072890, average: 0.266336, thin: 0.085998, skinny: 0.033565, fit: 0.222840, athletic: 0.213505, jacked: 0.00
7150
```

Predictive power of height, smokes, and how_active

Generating this model was much faster but not exactly more predictive.

For the “average” or “athletic” body categories, it does seem to do a better job, so the selected features may be more predictive for certain classifications than for others.

This does makes some sense: someone who reports an “athletic” body type is probably more likely to report physical activities in their essays.

Other correlations in the health data

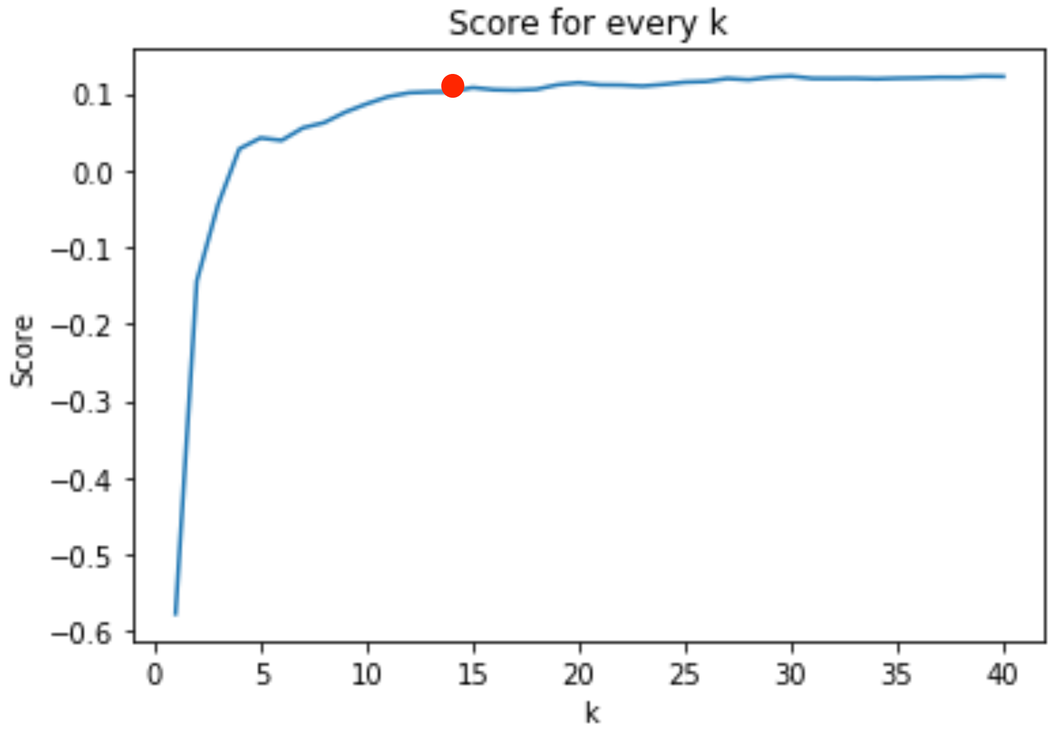
One of the things that is often speculated is that people who are more attractive tend to make more money.

I wanted to see if using self-reported body type as a proxy for attractiveness, a model could predict how much money someone makes.

To begin, I used body type and age to create a regressor.

Does body type and age predict income?

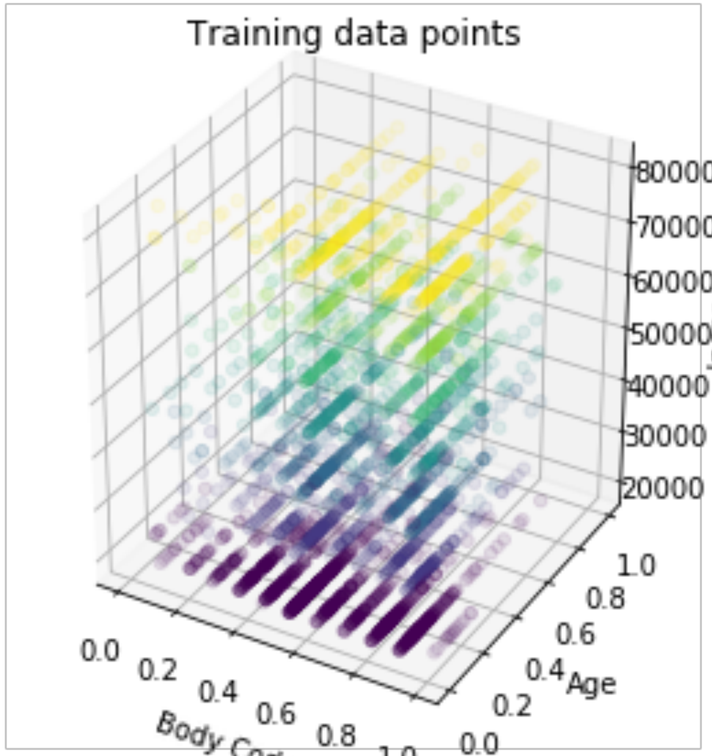
I tried predicting income using a K Nearest Neighbors Regressor

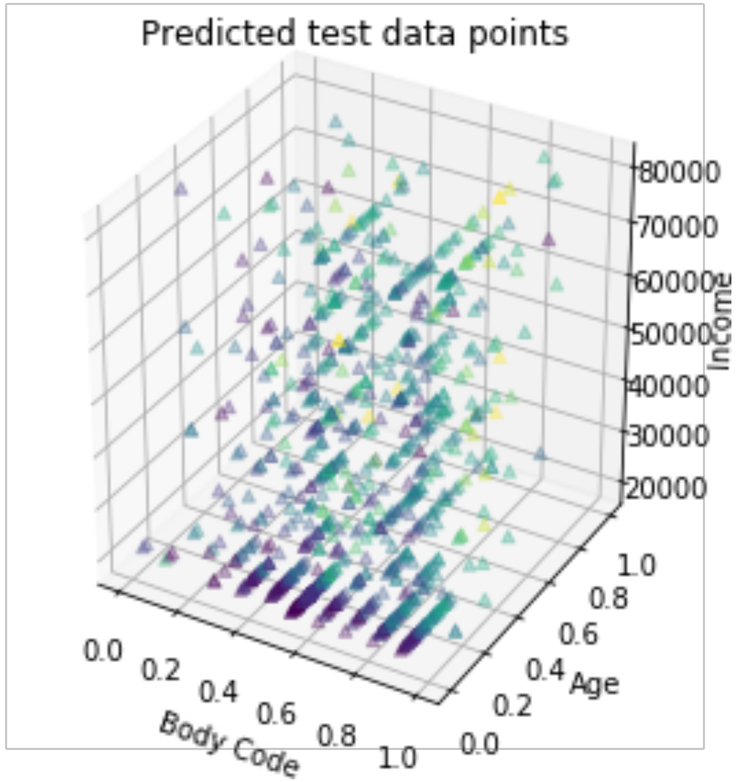


A test showed k of 29 to most accurate, but I chose 13 to minimize neighbor noise

Does body type and age predict income?

I tried predicting income using a K Nearest Neighbors Regressor (k = 13)

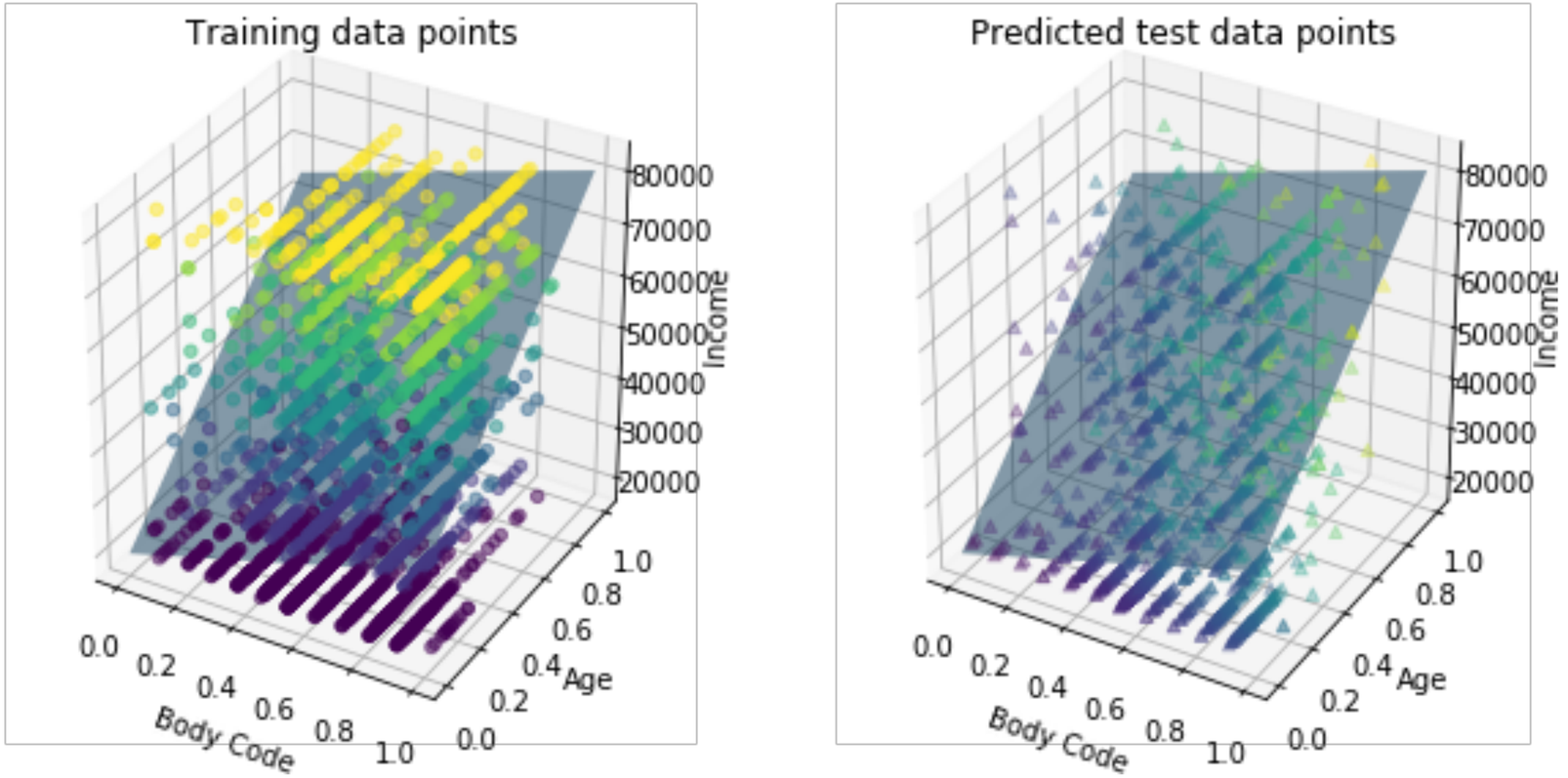




This regressor had a score of 0.1018 for this training set. Not great.

Does body type and age predict income?

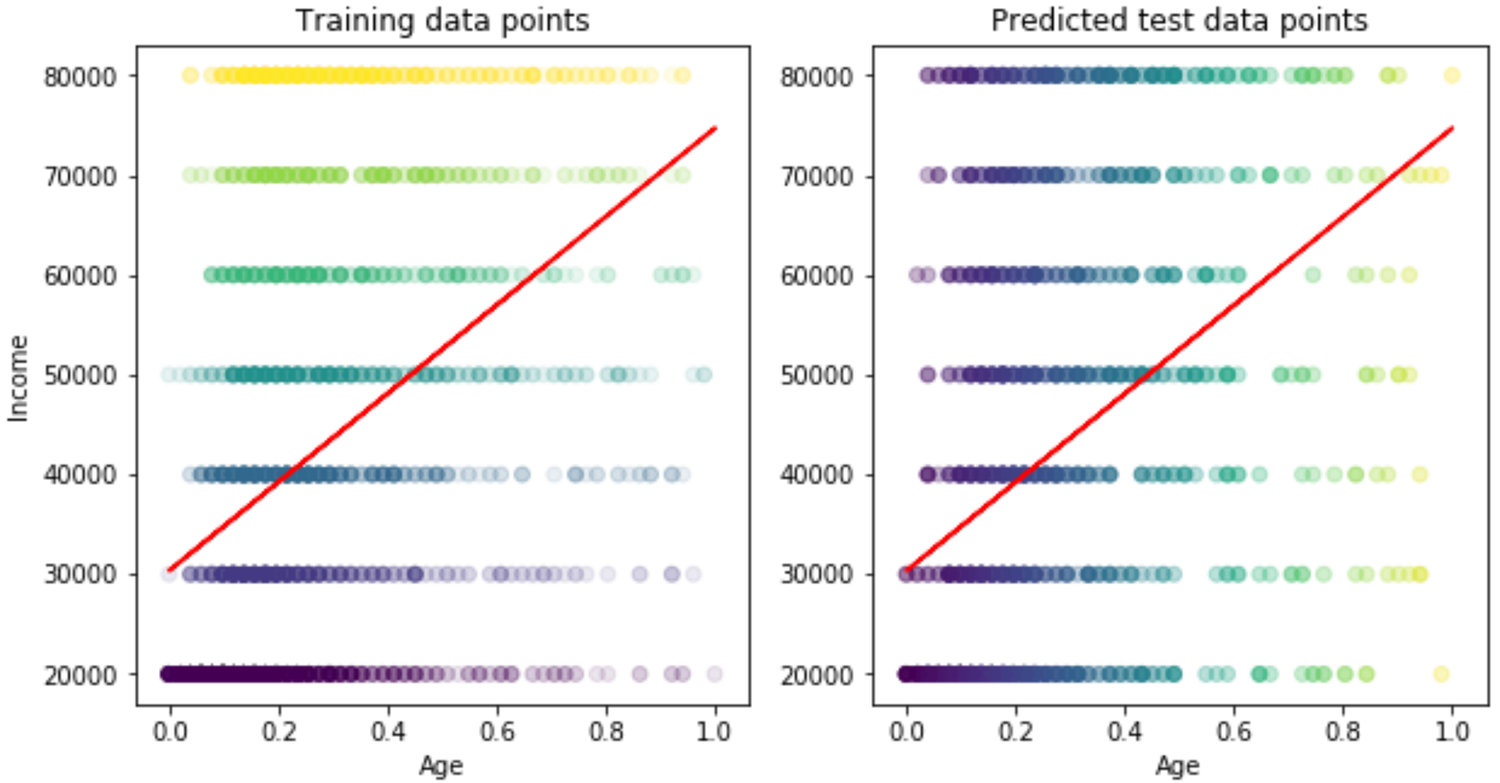
I next tried creating a linear regressor using LinearRegression()



This regressor had a score of 0.1556. Not great, but better than the KN model.

Does body type and age predict income?

I had a suspicion that age was the main determinant in this fit, so I used only age as a feature:



This regressor had a score of 0.1301. Slightly worse than including body type.

Things to note:

These data may reflect actual population demographics, but it's also possible that the self-reported nature of these features leads to an artificial bias.
For example: an "average" body type could mean any number of things, depending on location, or even whether someone feels comfortable being specific about their body type.

In the context of a dating questionnaire, people may be incentivized to skew what they report so as to not put off potential matches.

For future analyses:

Ideally, these analyses and models would have been run on real census and medical data (not self-reported) so that we can generate models more representative of the true population.