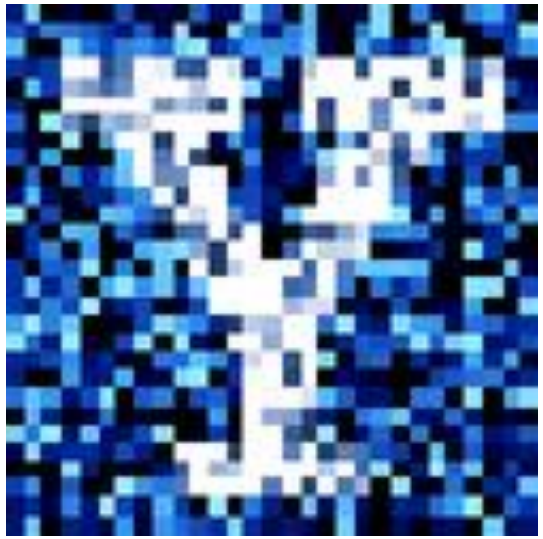
YData: Introduction to Data Science



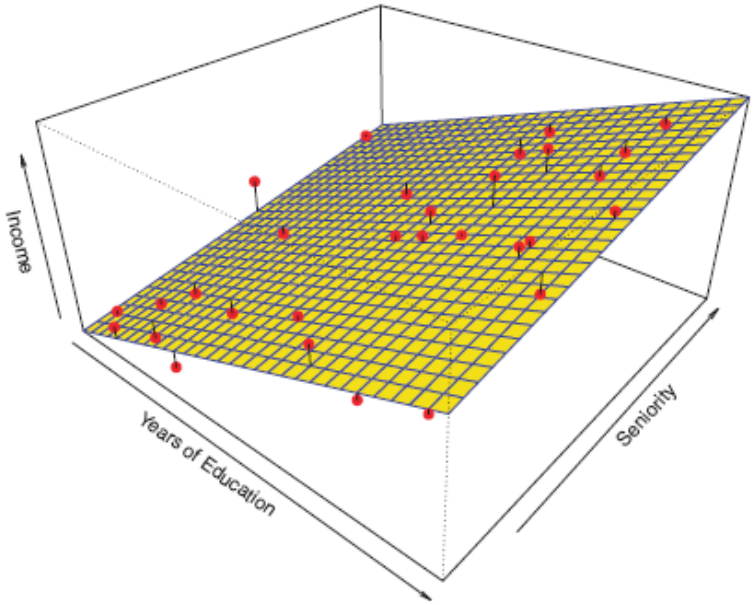
Lecture 38: pandas and conclusions

Overview

Multiple regression

pandas

Conclusions



pandas

Announcements

Homework 11 has been posted

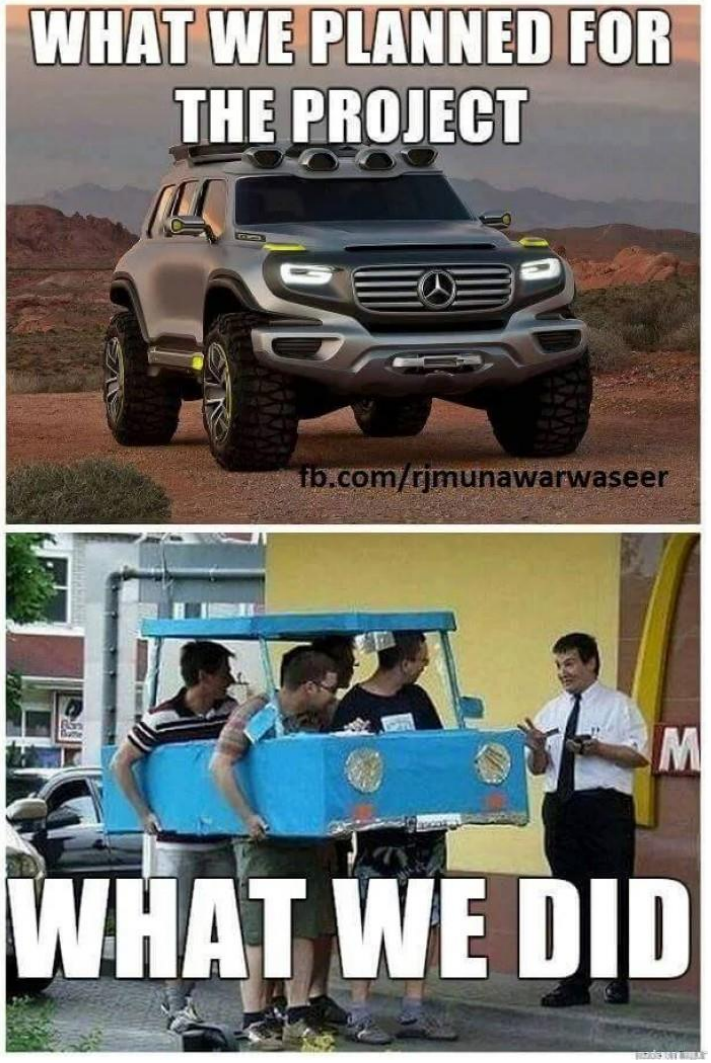
- It is due on Sunday May 1st

Project 3 is due tonight at 11pm

A practice final exam has been posted to Canvas

Final exam review session with be on Wednesday May 4th at 2:30pm

- In this classroom



Multiple regression

Prediction: regression and classification

We “learn” a function f

- f(**x**) ⟶ y

Input: **x** is a data vector of "features"

Output:

- Regression: output is a real number ($y \in R$)
- Classification: output is a categorical variable y_k

Multiple regression

In multiple regression we try to predict a quantitative response variable *y* using several features $x_1, x_2, \ldots, x_k$

We estimate coefficients using a data set to make predictions ŷ

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 \cdot x_1 + \hat{\beta}_2 \cdot x_2 + \ldots + \hat{\beta}_k \cdot x_k$$

Prediction: regression and classification

Potential Labels (y's)

Features (x's)

Classification labels (y's)

Regression labels (y's)

Bldg Type	SalePrice	Year Built	BsmtFin SF 1	1st Flr SF	2nd Flr SF	Gr Liv Area
1Fam	215000	1960	639	1656	0	1656
1Fam	189900	1997	791	928	701	1629
1Fam	195500	1998	602	926	678	1604
TwnhsE	213500	2001	616	1338	0	1338
TwnhsE	191500	1992	263	1280	0	1280
TwnhsE	236500	1995	1180	1616	0	1616
1Fam	189000	1999	0	1028	776	1804
1Fam	175900	1993	0	763	892	1655

Multiple regression

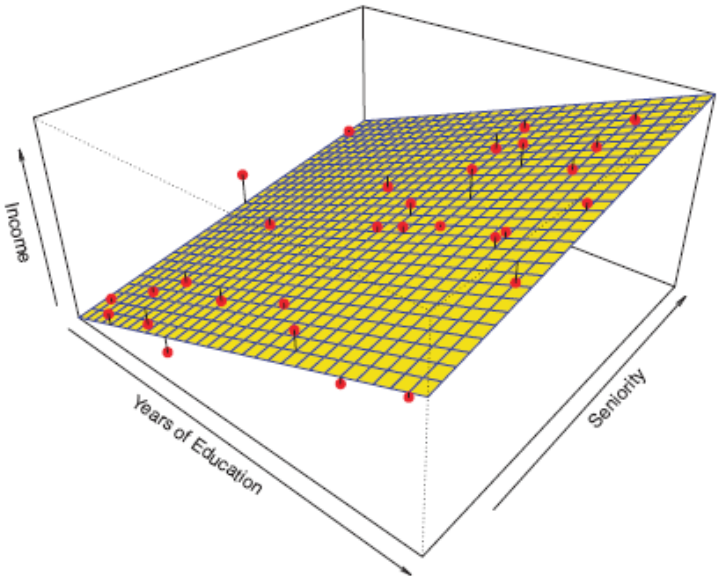
$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 \cdot x_1 + \hat{\beta}_1 \cdot x_2$$

$$\text{sales price} = \hat{\beta}_0 + \hat{\beta}_1 \cdot \text{square-footage} + \hat{\beta}_1 \cdot \text{year-built}$$

The coefficients ($\hat{\beta}_i$) are found by minimizing the RMSE

- i.e., we can use the minimize() function

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$



Let's explore this in Jupyter!

pandas

pandas

pandas is a software library written for the Python programming language for data manipulation and analysis

All the Table functions from Berekely's datascience package that we used in this class can be done using the padas package

- Some of the syntax can be more complicated
- But one can do more complex operations

Let's explore this in Jupyter!

Wrap up…