BIBA: Business Intelligence and Big Data

2018-10-03

Jens Ulrik Hansen

Correlation and Causation

Today's program

- Recap from last time
  - Feedback on the synopsis hand-in
- Correlation and causation
- Modeling
- Linear regression
- Marketing Mix Modeling (application of linear regression)

Recap from last time

- Plotting in R (ggplot2 package), the pipe operator "%>%" in R
- Exploratory Data Analysis
  - Understand your data
    - The format and data type of you variables
    - The mean of your variables
    - Missing values and outliers
  - Summarizing your data
    - Basic descriptive statistics: Counts, range, mean, median, variance, standard deviation, …
  - Visualizing the distributions of your variables
    - Histograms, boxplots, bar plots
  - Visualizing and quantifying the pairwise relationship between your variables
    - Boxplots, mosaic plots, scatterplots
    - Correlation coefficient

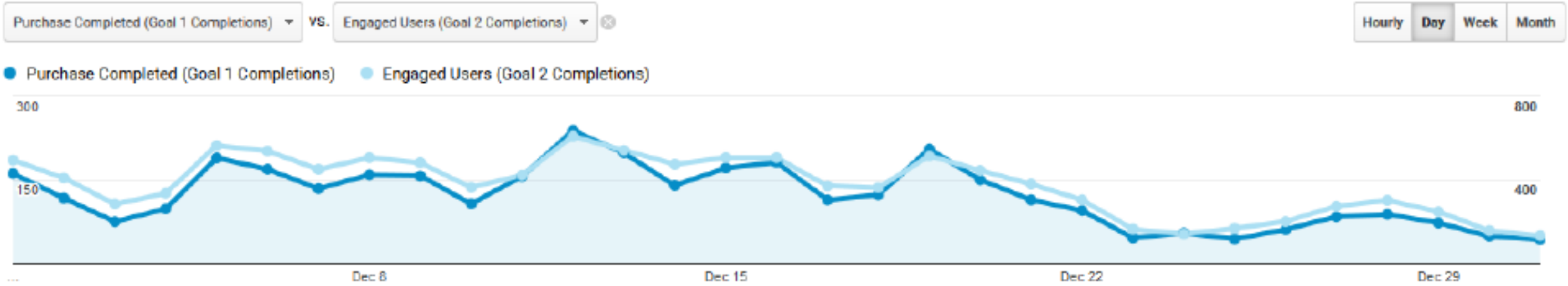
Feedback on the synopsis hand-in

- In general, good hand-ins
- The size of this hand-in varies a lot depending on
  - The complexity of the chosen data
  - The effort put into it
- I see a lot of different plots and different ways of creating the same plots
  - This is all good, it might just be because you have googled your way to a plot
- Please add a little bit of descriptive text
- The purpose of this hand-in was to get your hands dirty working with R and for you to get to know your data – for your own sake!

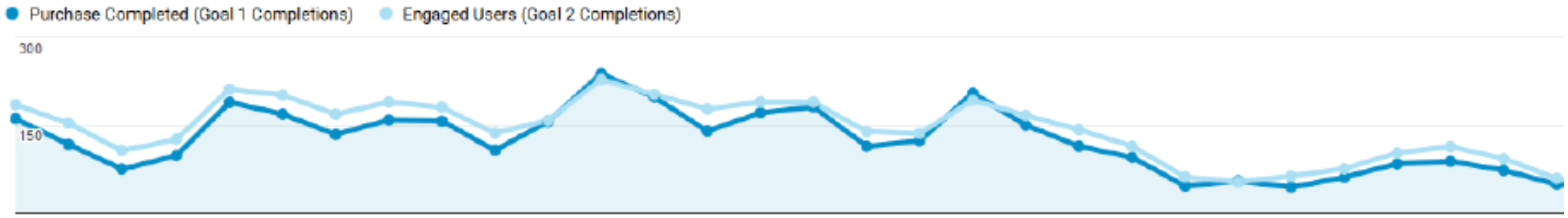
Correlation and causation

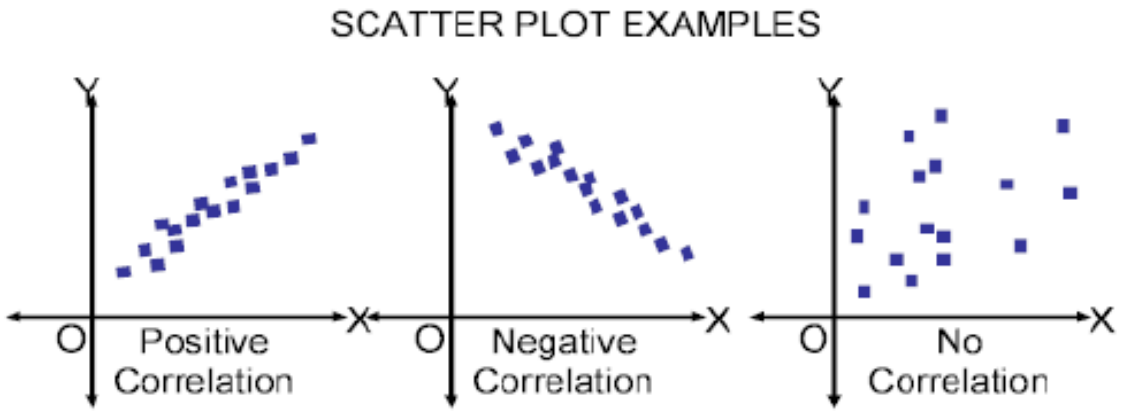
Correlation

- ***Correlation***: When two variables varies together, i.e.:
  - *There is a tendency that the second goes up when the first one goes up, and there is a tendency that the second goes down when the first one goes down (positive correlation)*
  - *There is a tendency that the second goes down when the first one goes up, and there is a tendency that the second goes up when the first one goes down (negative correlation)*
  - Examples of correlation: height and weight, engagement and sales:



How to detect correlation

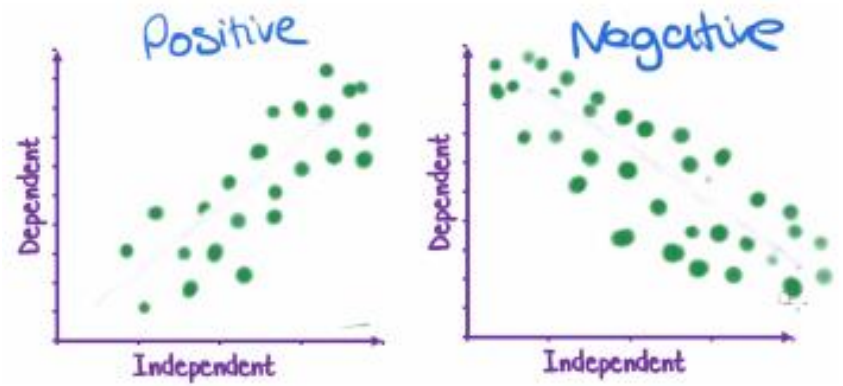


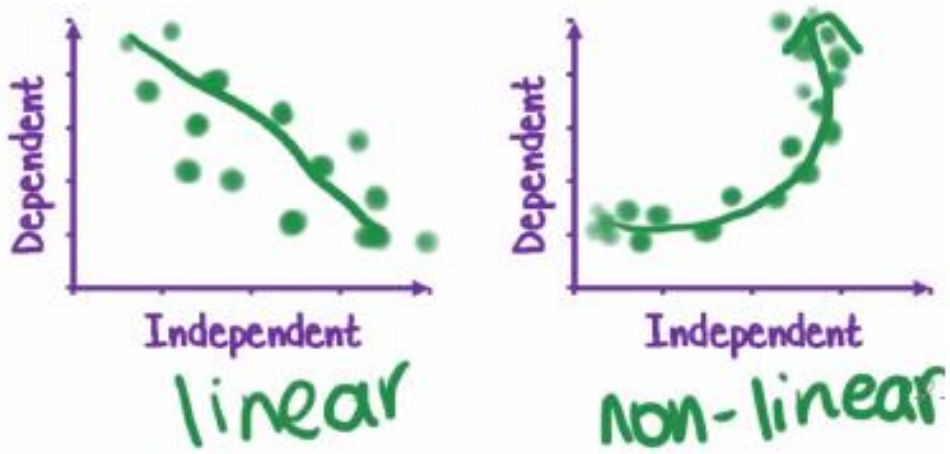


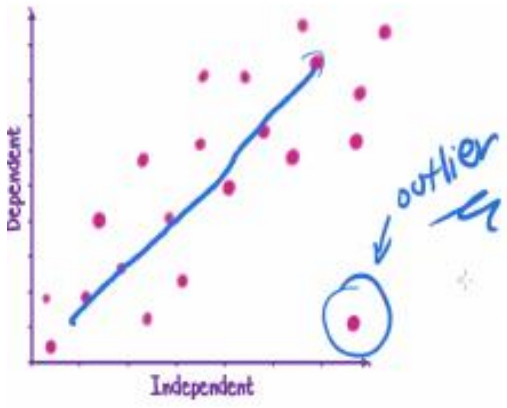
- Visualizing correlation?
  - Scatter plots/x-y-plots
- Quantifying correlation?
  - (Pearson's) Correlation coefficient
    - A number between -1 and 1. 1 is perfect positive correlation, -1 is perfect negative correlation, and 0 is no correlation.
    - Only quantifies linear correlation

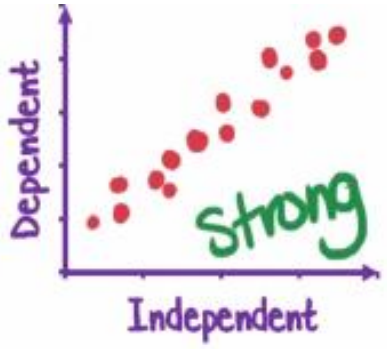
Types of correlations (in scatter plots)

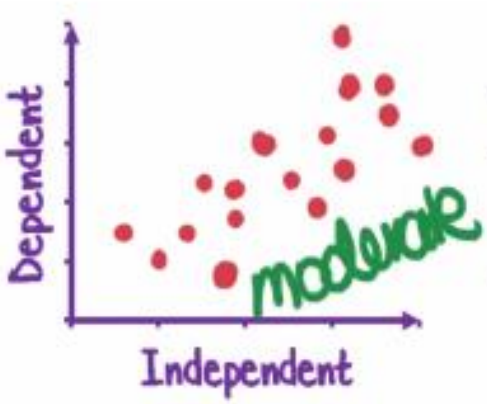
- See: https://www.youtube.com/watch?v=PE_BpXTyKCE
- **Direction**
  - Positive
  - negative
- **Shape**
  - Linear
  - non-linear
- **Strength**
  - Weak
  - Moderate
  - strong
- **Outliers**

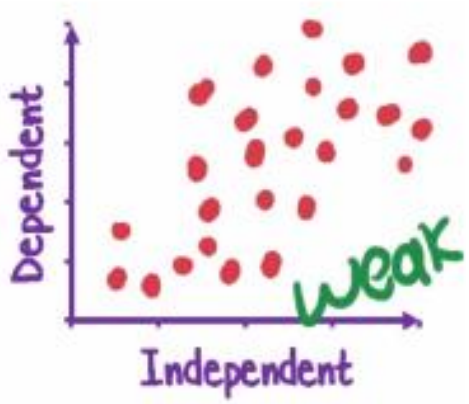






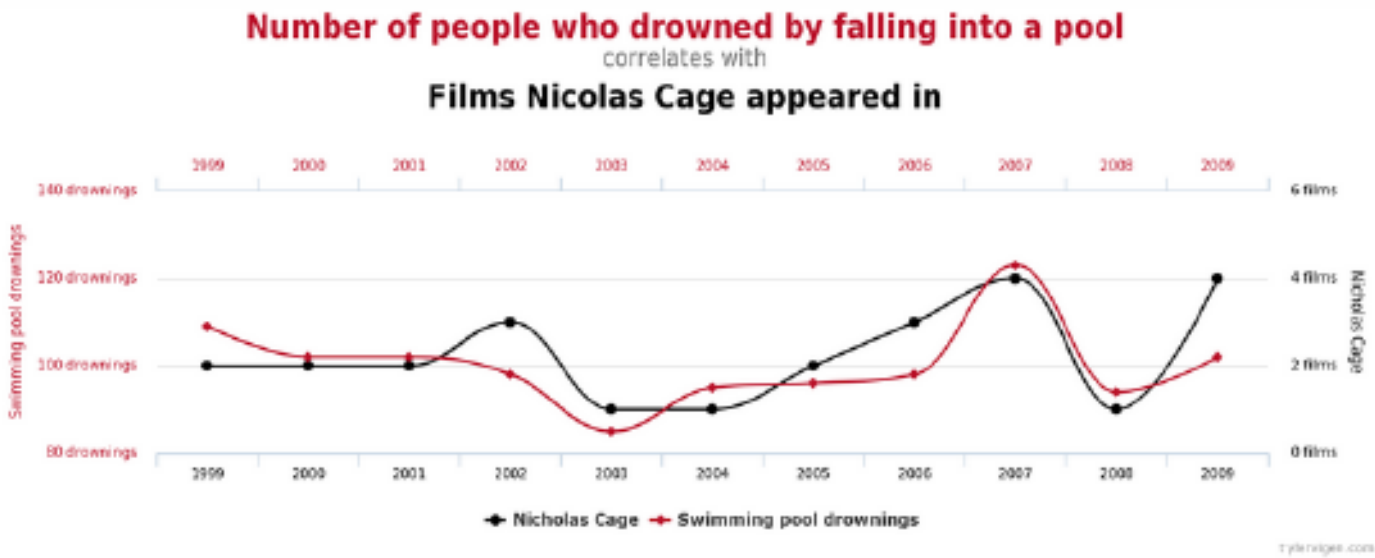


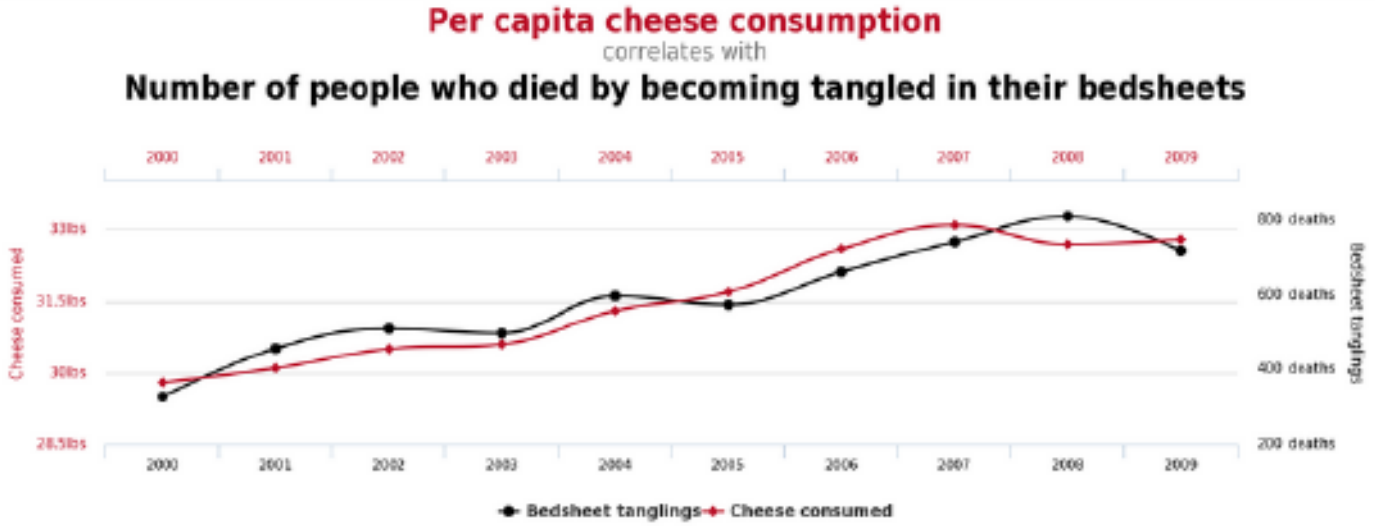




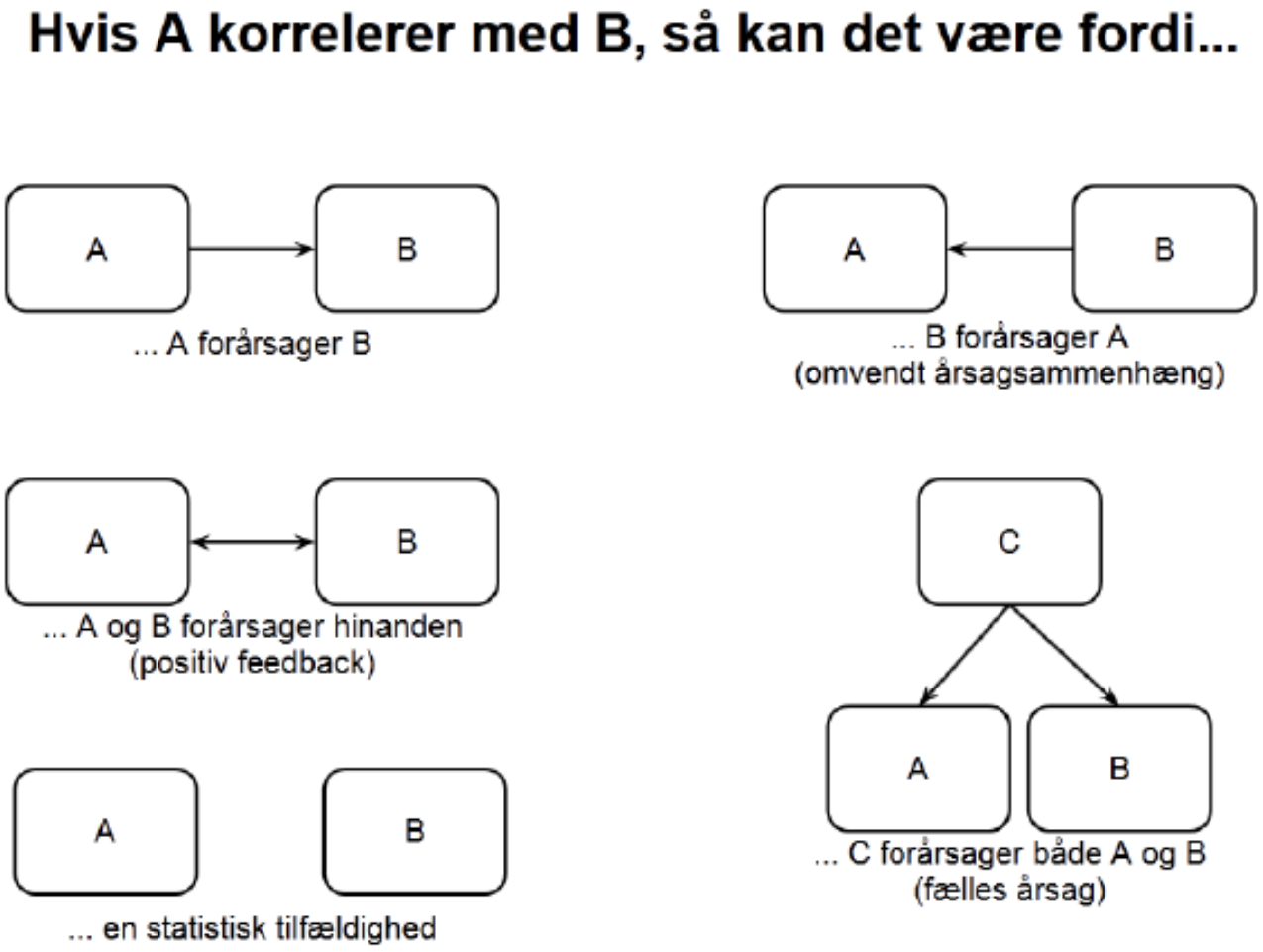
Correlation and causation

- ***Just because two variables correlates, it does not mean that there is a causal relationship between them***
- Spurious Correlations
  - (http://www.tylervigen.com/spurious-correlations)
- There can be multiple explanations...





Correlation and causation



- If A correlates with B it can be because of…
- Examples:
  - Grades in high-school and in college
  - Rain and sun
  - Height and weight
  - Engaged users and Purchased completed

Why causation?

- If X causes Y
  - We may be able to predict Y based on X
    - If the weather causes our sales of ice cream, we can predict our ice cream sales if we know the weather forecast (we are not in control of the weather)
  - We can “control” Y by manipulating X
    - If our unhappy costumers are caused by the long waiting time in our customer service, we can lover the waiting time in our customer service and thereby decrease happiness
    - If marketing spend causes sales numbers, we can increase our sales by increasing our marketing spend
- If X and Y just correlates
  - We might still be able to predict Y based on X,
  - but we might not be able to control Y by manipulating X

Correlation and predictions

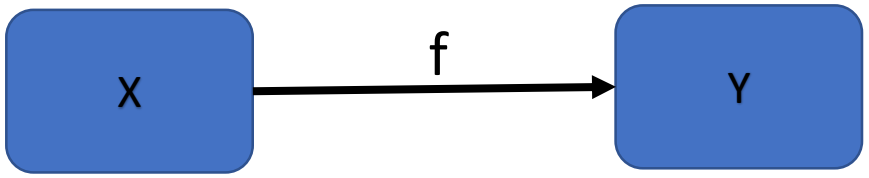
- Can we find the a formula for the line representing correlation?
  - Yes, that's is exactly what linear regression is about!
- Can this approach be generalized?
  - Yes, we can talk about predictive modeling in general, or statistical modeling or machine learning modeling
  - We can do all sorts of predictive modeling beyond linear regression of one variable (Y) on one other variable (X)

Correlation exercises

- Please do the correlation exercises in the notebook “Correlation exercises.ipynb” in the BIBA-2018 Library on Azure notebooks.

Modeling

Predictive modeling



- A model in general
  - A function that takes a variable value as input (a value of X) and produces a value of another variable (Y) : f(X) = Y
  - If we have such a function, we can always predict Y from any value of Y (this is the definition of a function!)
  - There can be more than one input variable: f(X_1, X_2, …, X_n) = Y
- Business examples of predictions
  - Predicting house prices
  - Predicting sales or new costumers based on media etc.
  - Predicting churn
  - Predicting retention of employees
  - Predicting credit score

Modeling in general

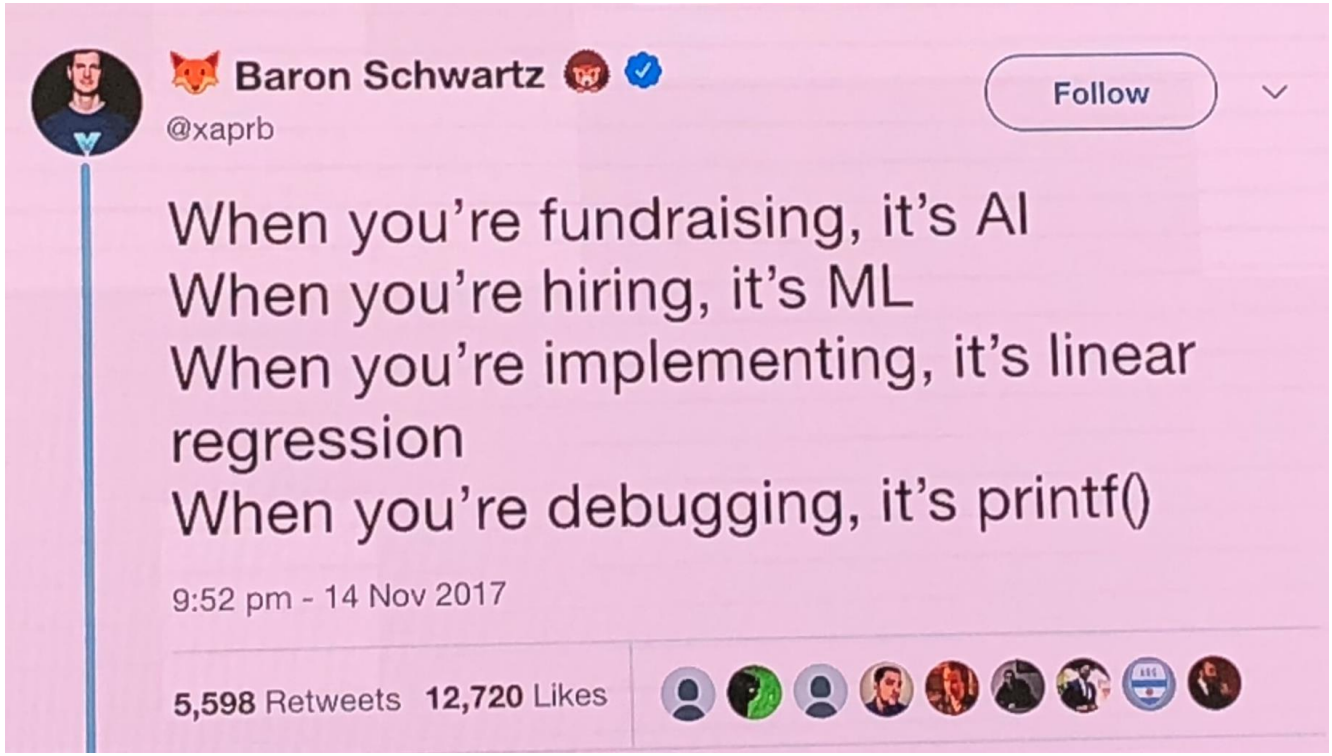
- Modeling is about building models (functions f such that f(X)=Y)
- However modeling in a very broad sense is:
  - As soon as you try to describe a natural phenomenon with mathematics
  - As soon as you try to conceptualize a “natural phenomenon” in a structured way
- Modeling can serve multiple purposes
  - Provide insight into particular phenomena or data
  - Discover patters in data
  - Predict future effects and event
  - . . . etc.
- Models partition data into patterns and residuals

Three steps to modeling

1. Define ***a family of models***. A class of models you want to consider for your data.
   - Example, y = a + b * x_1 + c * x_2 (a, b, and c are unknown constants/parameters)
   - Each specific combination of values for a, b, and c give rise to one model
2. ***Fit a model to data***. Find the model from the family that is closest to your data
   - Example: find specific values for a, b, and c that make you model (y = a * x_1 + b * x_2 + c) "closest" to your data
   - What does "closest" mean? (one option is "root-mean-square-error")
3. ***Evaluate your model***
   - Just because you have found the model closest to your data coming from your family of models, does not mean that your model is true or even good.
   - You could have chosen the wrong family of models to start with (y might not be a linear function of x_1 and x_2)
   - Your data could be so noisy that it is hard to judge whether a model is close to it
   - Thus, it is important to proper evaluate your model
     - However: "all models are wrong, but some are useful" (George Box)

Linear regression

Linear regression



- Linear regression is the easiest regression model to use and understand
- Linear regression (and its extensions) is good enough for many business intelligence problems
- Linear regression models and predictions based on them can be explained and easily communicated
- Linear regression provides a baseline to which more advanced and sophisticated prediction modeling techniques can be compared

Linear regression

- See the notebook: “Linear Regression.ipynb”

Marketing Mix Modelling

Marketing Mix Modelling

- See the notebook: “Marketing Mix Modeling.ipynb”

Exercise

- Try to fit a linear regression to some of you data for your business case