SUMMER UNIVERSITY

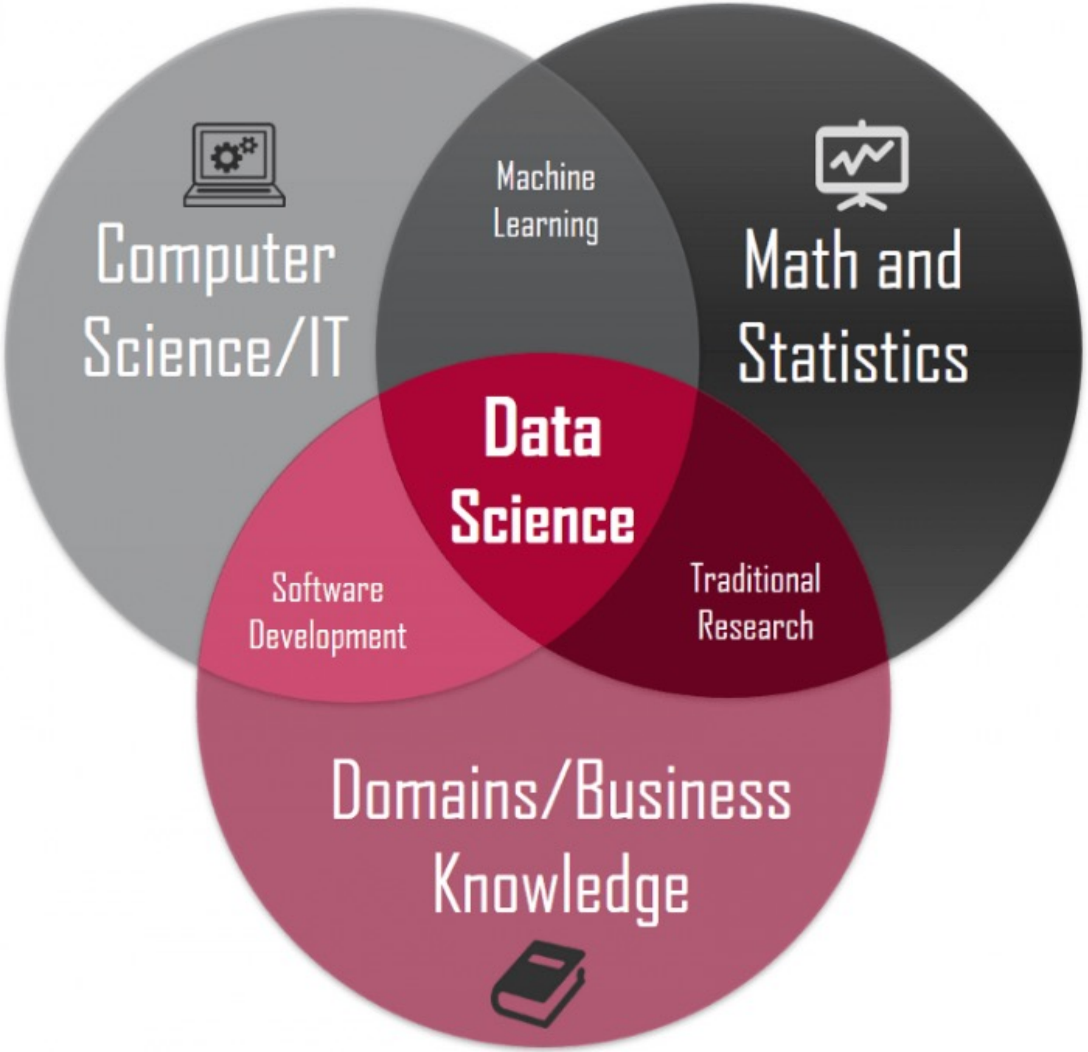
Technische Universität Berlin

Advanced Data Science

Dr. Hamid Mostofi

mostofidarbani@tu-berlin.de

Data science



Python Libraries for Data Science

Many popular Python toolboxes/libraries:

- NumPy
- SciPy
- Pandas
- SciKit-Learn

Visualization libraries

- matplotlib
- Seaborn

and many more …

Python Libraries for Data Science

NumPy:

- introduces objects for multidimensional arrays and matrices, as well as functions that allow to easily **perform advanced mathematical** and statistical operations on those objects

- provides vectorization of mathematical operations on arrays and matrices which significantly improves the performance

- many other python libraries are built on NumPy

Link: http://www.numpy.org/

SciPy.org

Python Libraries for Data Science

SciPy:

- collection of algorithms for linear algebra, differential equations, numerical integration, optimization, statistics and more
- part of SciPy Stack
- built on NumPy

Link: https://www.scipy.org/scipylib/

pandas

$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$

Python Libraries for Data Science

Pandas:

- adds data structures and tools designed to work with table-like data (similar to Series and Data Frames in R)

- provides tools for data manipulation: reshaping, merging, sorting, slicing, aggregation etc.

- allows handling missing data

Link: http://pandas.pydata.org/

Python Libraries for Data Science

SciKit-Learn:

- provides machine learning algorithms: classification, regression, clustering, model validation etc.

- built on NumPy, SciPy and matplotlib

Link: http://scikit-learn.org/

matplotlib:

- python 2D plotting library which produces publication quality figures in a variety of hardcopy formats
- a set of functionalities similar to those of MATLAB
- line plots, scatter plots, barcharts, histograms, pie charts etc.
- relatively low-level; some effort needed to create advanced visualization

Link: https://matplotlib.org/

Seaborn:

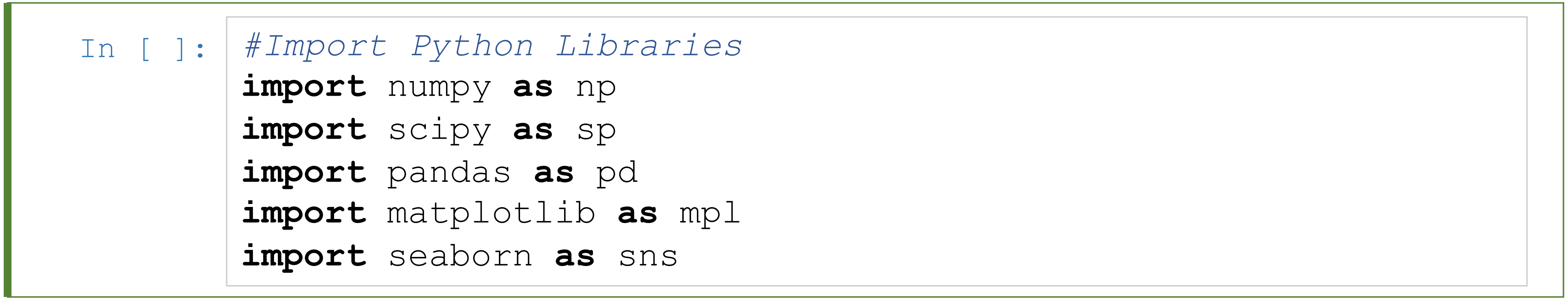
- based on matplotlib

- provides high level interface for drawing attractive statistical graphics

- Similar (in style) to the popular ggplot2 library in R

Link: https://seaborn.pydata.org/

Loading Python Libraries

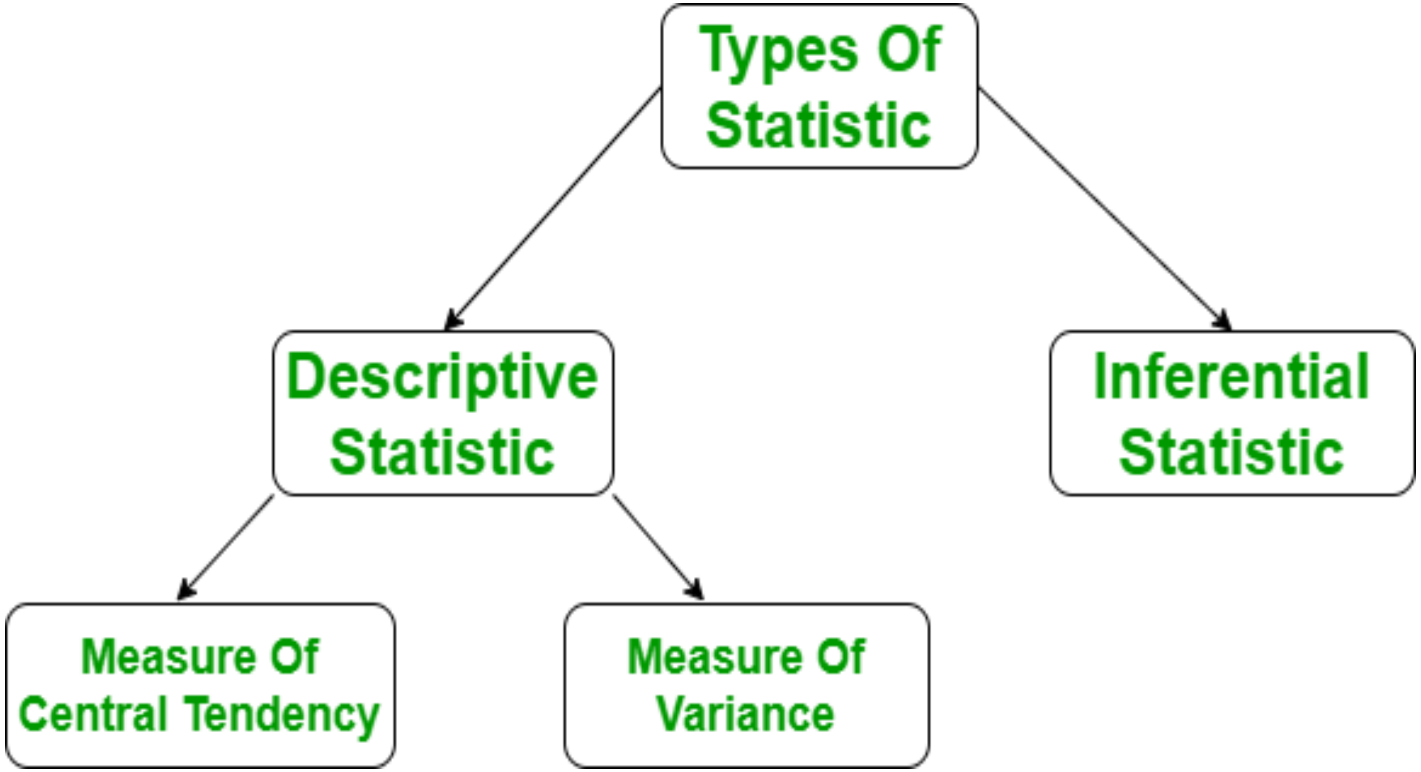


```
#Import Python Libraries
import numpy as np
import scipy as sp
import pandas as pd
import matplotlib as mpl
import seaborn as sns
```

Press `Shift+Enter` to execute the *jupyter* cell

What we learned

- To compare the data of the two or more groups together, we should consider the measures of **Central Tendency** and **Dispersion** together.



One-Way ANOVA in Python:

- One-way ANOVA (also known as “analysis of variance”) is a test that is used to find out whether there exists a statistically significant difference between the mean values of more than one group.
- **Hypothesis involved:**
- A one-way ANOVA has the below given null and alternative hypotheses:
- H0 (null hypothesis): μ1 = μ2 = μ3 = … = μk (It implies that the means of all the population are equal)
- H1 (null hypothesis): It states that there will be at least one population mean that differs from the rest

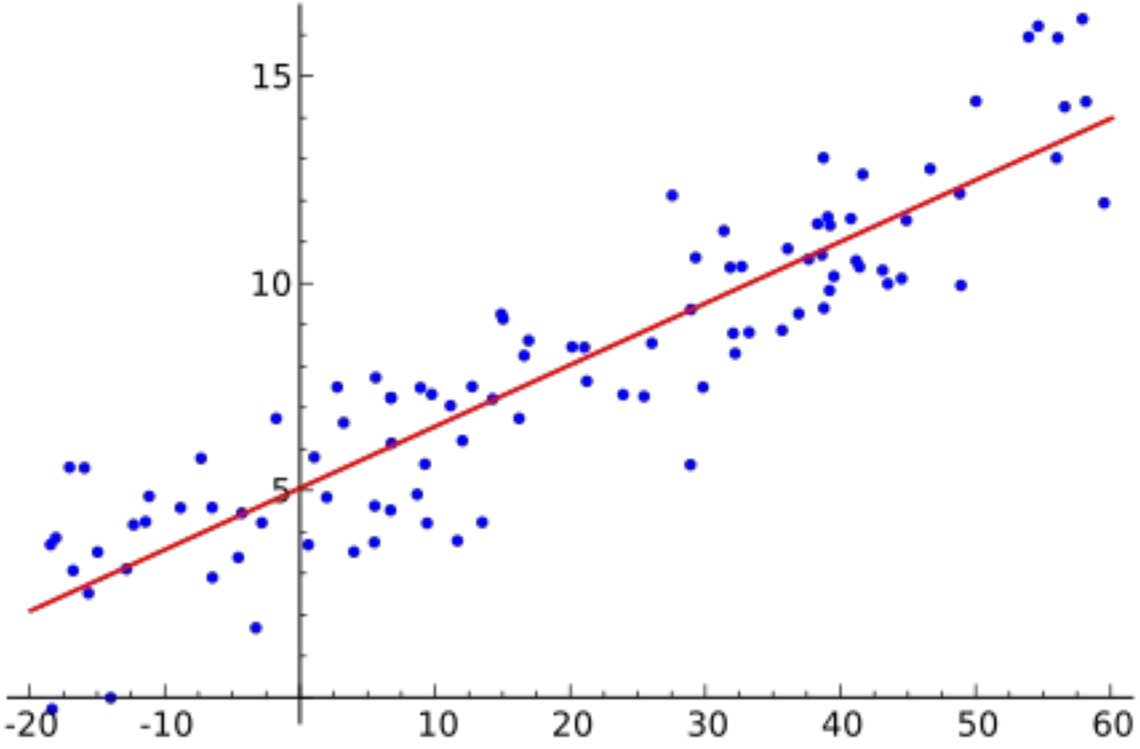
Statement

Researchers took 20 cars of the same to take part in a study. These cars are randomly doped with one of the four-engine oils and allowed to run freely for 100 kilometers each. At the end of the journey, the performance of each of the cars is noted. **Which of them is better?**

Before proceeding further we need to install the SciPy library in our system. You can install this library by using the below command in the terminal:

What is Regression?

- **Linear regression** is a model that predicts a relationship of direct proportionality between the dependent variable (plotted on the vertical or Y axis) and the predictor variables (plotted on the X axis) that produces a straight line, like so:



Predicting Housing Prices with Linear Regression

- For our dependent variable we'll use housing_price_index (HPI), which measures price changes of residential housing.

- For our predictor variables, we use our intuition to select drivers of macro- (or “big picture”) economic activity, such as unemployment, interest rates, and gross domestic product (total productivity)

Simple Linear Regression

- Simple linear regression uses a single predictor variable to explain a dependent variable.

  A simple linear regression equation is as follows:

$$Y_i = \alpha + \beta x_i + \epsilon_i$$

- Where:
- y = dependent variable
- β = regression coefficient
- α = intercept (expected mean value of housing prices when our independent variable is zero)
- x = predictor (or independent) variable used to predict Y
- ϵ = the error term, which accounts for the randomness or other unknown variables that our model can't explain.

Statsmodels’ OLS function,

- Using statsmodels' ols function, we construct our model setting housing_price_index as a function of total_unemployed.
- We assume that an increase in the total number of unemployed people will have downward pressure on housing prices.
- Maybe we're wrong, but we have to start somewhere!

- **Adj. R-squared** indicates that 95% of housing prices can be explained by our predictor variable, total_unemployed.

- The **regression coefficient** **(coef)** represents the change in the dependent variable resulting from a one unit change in the predictor variable, all other variables being held constant. In our model, a one unit increase in total_unemployed reduces housing_price_index by 8.33. In line with our assumptions, an increase in unemployment appears to reduce housing prices.

- The **standard error** measures the accuracy of total_unemployed's coefficient by estimating the variation of the coefficient if the same test were run on a different sample of our population. Our standard error, 0.41, is low and therefore appears accurate.

- The **p-value** means the probability of an 8.33 decrease in housing_price_index due to a one unit increase in total_unemployed is 0%, assuming there is no relationship between the two variables. A low p-value indicates that the results are statistically significant, that is in general the p-value is less than 0.05.
- The **confidence interval** is a range within which our coefficient is likely to fall. We can be 95% confident that total_unemployed's coefficient will be within our confidence interval, [-9.185, -7.480].

Confidence Intervals for Regression Coefficients

- A 95% confidence interval for βi has two equivalent definitions:

The interval is the set of values for which a hypothesis test to the level of 5% cannot be rejected. The interval has a probability of 95% to contain the true value of βi.

So in 95% of all samples that could be drawn, the confidence interval will cover the true value of βi.

We also say that the interval has a confidence level of 95%.

- **A Confidence Interval for βiImagine you could draw all possible random samples of given size. The interval that contains the true value** βi **in** 95% **of all samples is given by the expression** $CI\beta_i 0.95=[\beta i-1.96\times SE(\beta i),\ \beta i+1.96\times SE(\beta i)]$.
- **Equivalently, this interval can be seen as the set of null hypotheses for which a** 5% **two-sided hypothesis test does not reject.**

- The standard error of the coefficient measures how precisely the model estimates the coefficient's unknown value. The standard error of the coefficient is always positive.

Multiple Linear Regression

- Mathematically, multiple linear regression is:

$$Y=\beta_0+\beta_1x_1+\beta_2x_2+...+\beta_kx_k+\epsilon$$

We know that unemployment cannot entirely explain housing prices. To get a clearer picture of what influences housing prices, we add and test different variables and analyze the regression results to see which combinations of predictor variables satisfy **OLS** assumptions, while remaining intuitively appealing from an economic perspective.

- We assume at a model that contains the following variables:

fed_funds, consumer_price_index, long_interest_rate, and gross_domestic_product, in addition to our original predictor, total_unemployed.

- Adding the new variables decreased the impact of total_unemployed on housing_price_index. total_unemployed‘s impact is now more unpredictable (**standard error** increased from 0.41 to 2.399), and, since the **p-value** is higher (from 0 to 0.943), therefore there is no significant relation by considering fed_funds, consumer_price_index, long_interest_rate, and gross_domestic_product,
-

- Although total_unemployed may be correlated with housing_price_index, our other predictors seem to capture more of the variation in housing prices.
- The real-world interconnectivity among our variables can't be encapsulated by a simple linear regression alone; a more robust model is required. This is why our multiple linear regression model's results change drastically when introducing new variables.
- That all our newly introduced variables are statistically significant at the 5% threshold, and that our coefficients follow our assumptions, indicates that our multiple linear regression model is better than our simple linear model.

Binary Logistic Regression in Python

- Binary logistic regression models the relationship between a set of independent variables and a binary dependent variable.
- It is useful when the dependent variable is dichotomous in nature, such as death or survival, absence or presence, pass or fail, for example. In logistic regression, the dependent variable is a binary variable that contains data coded as 1 (yes, success, etc.) or 0 (no, failure, etc.).
- In other words, the logistic regression model predicts P (Y=1)/ P (Y=0)as a function of X.
- Independent variables can be categorical or continuous, for example, gender, age, income, geographical region and so on.

Statistical Model – For k Predictors

- So what does the statistical model in binary logistic regression look like? In this equation, p is the probability that Y equals one given X, where Y is the dependent variable and X’s are independent variables. B 0 to B K are the parameters of the model.

$$\log\left(\frac{\mathrm{p}}{1-\mathrm{p}}\right) = \mathrm{b}_0 + \mathrm{b}_1\mathrm{X}_1 + \cdots + \mathrm{b}_\mathrm{k}\mathrm{X}_\mathrm{k}$$

- **where,**
- **p : Probability that Y=1 given X**
- **Y : Dependent Variable**
- **X1, X2 ,…, Xk : Independent Variables**
- **b0, b1 ,…, bk : Parameters of Model**

Case Study :

Modeling Loan Defaults

- One bank has the demographic and transactional data of its loan customers.
- It wants to develop a model that predicts defaulters and help the bank in its loan disbursal decision making.
- The objective here is to predict whether customers applying for a loan will be defaulters or not.
- The independent variables are age group, years at current address, years at current employer, debt to income ratio, credit card debt and other debt.
- All of these variables are collected at the time of the loan application process and will be used as independent variables. The dependent variable is the status observed after the loan is disbursed, which will be one if it is a defaulter and zero if not.

Background

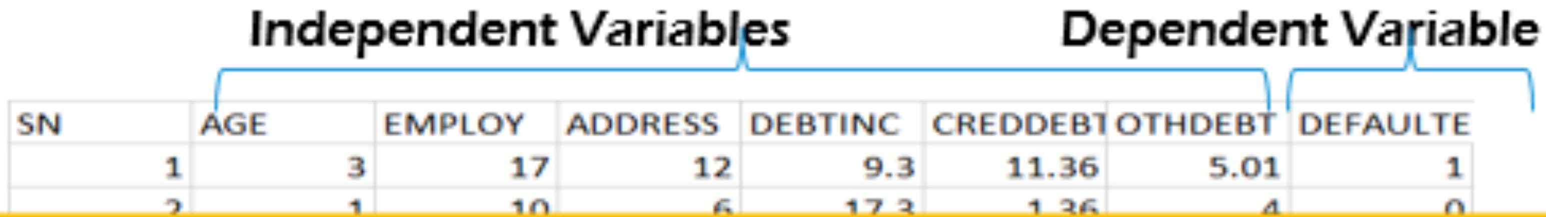
- A bank possesses demographic and transactional data of its loan customers. If the bank has a model to predict defaulters it can help in loan disbursal decision making.

Objective

- To predict whether the customer applying for the loan will be a defaulter or not.

Available Information

- **Sample size is 700**
- **Independent Variables**: Age group, Years at current address, Years at current employer, Debt to Income Ratio, Credit Card Debts, Other Debts. The information on predictors was collected at the time of loan application process.
- **Dependent Variable**: Defaulter (=1 if defaulter ,0 otherwise). The status is observed after loan is disbursed.



Independent Variables

Dependent Variable

SN	AGE	EMPLOY	ADDRESS	DEBTINC	CREDDEBT	OTHDEBT	DEFAULTE
1	3	17	12	9.3	11.36	5.01	1
2	1	10	6	17.3	1.36	4	0

Column	Description	Type	Measurement	Possible Values
SN	Serial Number	numeric	-	-
AGE	Age Groups	Categorical	1(<28 years), 2(28-40 years), 3(>40 years)	3
EMPLOY	Number of years customer working at current employer	Continuous	-	Positive value
ADDRESS	Number of years customer staying at current address	Continuous	-	Positive value
DEBTINC	Debt to Income Ratio	Continuous	-	Positive value
CREDDEBT	Credit to Debit Ratio	Continuous	-	Positive value
OTHDEBT	Other Debt	Continuous	-	Positive value
DEFAULTER	Whether customer defaulted on loan	Binary	1(Defaulter), 0(Non-Defaulter)	2