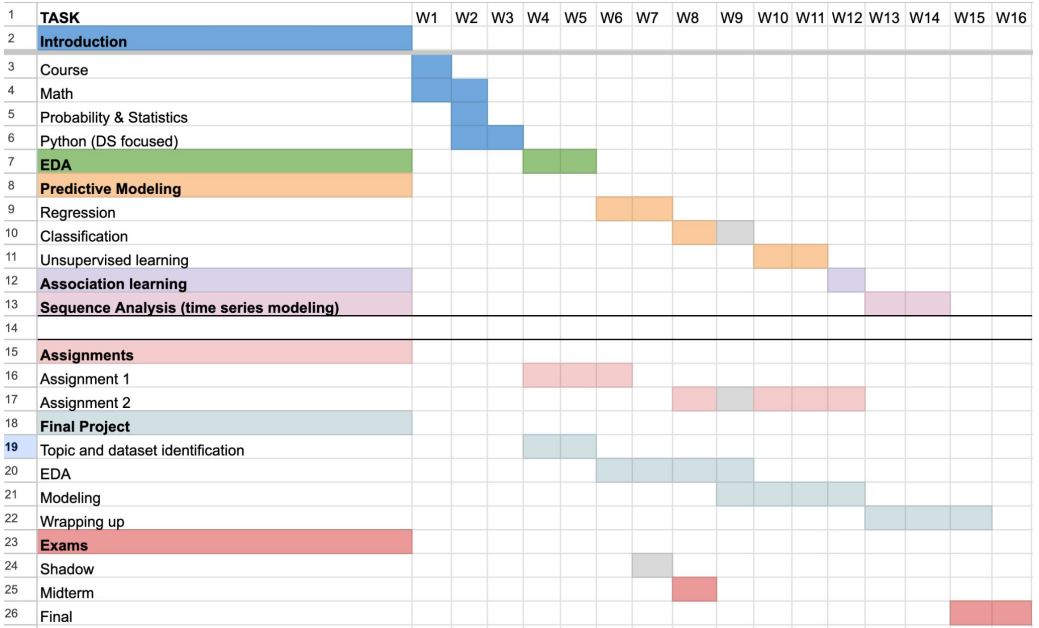
CIS 635 Knowledge Discovery & Data Mining

Course Review Week

Our Plan



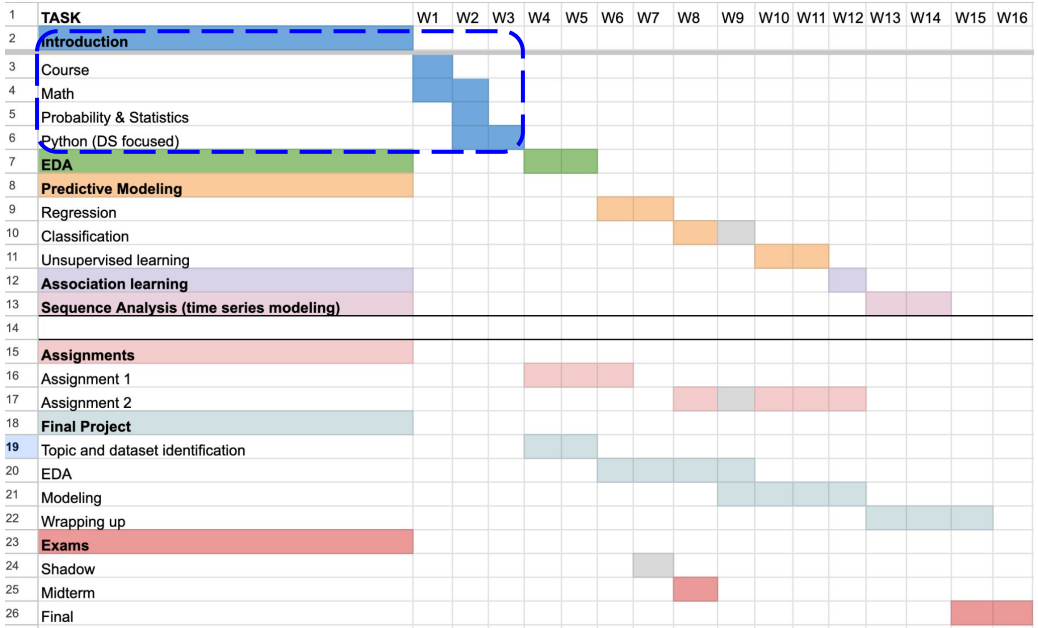
	TASK	W1	W2	W3	W4	W5	W6	W7	W8	W9	W10	W11	W12	W13	W14	W15	W16
2	**Introduction**																
3	Course																
4	Math																
5	Probability & Statistics																
6	Python (DS focused)																
7	**EDA**																
8	**Predictive Modeling**																
9	Regression																
10	Classification																
11	Unsupervised learning																
12	**Association learning**																
13	**Sequence Analysis (time series modeling)**																
14																	
15	**Assignments**																
16	Assignment 1																
17	Assignment 2																
18	**Final Project**																
19	Topic and dataset identification																
20	EDA																
21	Modeling																
22	Wrapping up																
23	**Exams**																
24	Shadow																
25	Midterm																
26	Final																

Background diagnostics + Python (DS focused)

Vector space
Proximity or distance metric
- L1/Manhattan distance
- L2/Euclidean distance,
- Cosine distances

kNN model:
- Distance based
- Can be applied to both Regression and Classification tasks

Probability (measuring uncertainty)
Probability distributions



	TASK	W1	W2	W3	W4	W5	W6	W7	W8	W9	W10	W11	W12	W13	W14	W15	W16
2	**Introduction**																
3	Course																
4	Math																
5	Probability & Statistics																
6	Python (DS focused)																
7	**EDA**																
8	**Predictive Modeling**																
9	Regression																
10	Classification																
11	Unsupervised learning																
12	**Association learning**																
13	**Sequence Analysis (time series modeling)**																
14																	
15	**Assignments**																
16	Assignment 1																
17	Assignment 2																
18	**Final Project**																
19	Topic and dataset identification																
20	EDA																
21	Modeling																
22	Wrapping up																
23	**Exams**																
24	Shadow																
25	Midterm																
26	Final																

Introduction to Data Mining + KDD Process

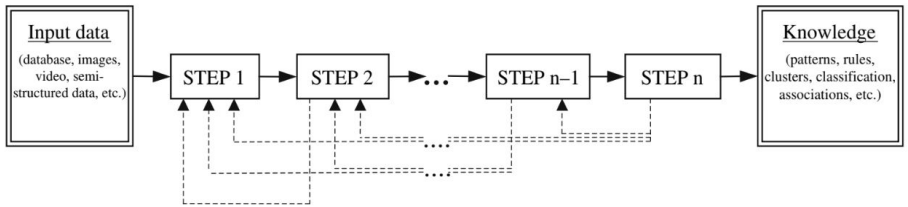
- Process of **extracting and discovering patterns in large datasets**
- Involves methods: **ML, Statistics, DBMS**
- Interdisciplinary field: **CS, Statistics**
- Overall goal:
    - Extracting information from dataset
    - Transform into a comprehensive structure for further use
- Data mining is the **analysis step of**
    - The **KDD**
    - Aside from raw analysis, it also involves
        - Database and data management aspects
        - Data preprocessing
        - Modeling and inference considerations
        - Evaluation and metrics
        - Post processing of discovered structures and visualizations

Wikipedia

Knowledge Discovery Process (KDP) Models

Academic Research Models

- Introduced in the mid 1990s
- Several models available
- Suggested steps are similar



9 steps: Fayyard et al KDP model:

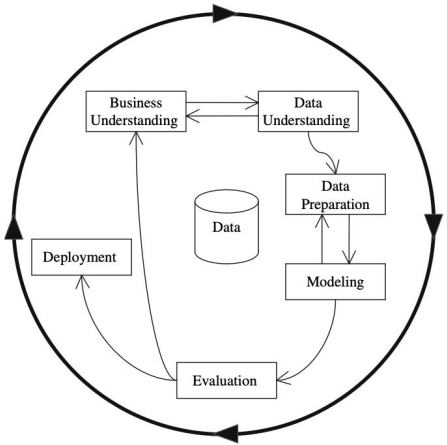
1) Understanding the application domain
2) Creating a target dataset
3) Data cleaning and preprocessing
4) Data reduction and projection
5) Choosing the data mining task
6) Choosing the algorithm
7) Data mining
8) Interpreting mined patterns
9) Consolidating discovered patterns

Knowledge Discovery Process (KDP) Models

Industrial Models

- Business understanding
- Data Understanding
- Data preparation
- Modeling
- Evaluation
- Deployment

Data Mining - A Knowledge Discovery Approach by Cis Pedrycz, and Swiniarski



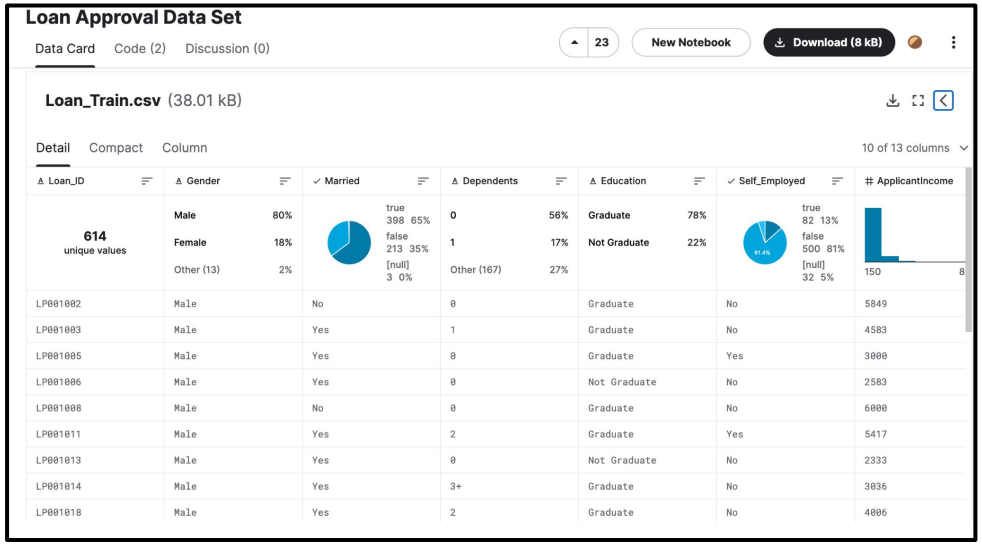
The CRISP-DM KD process model (source: http://www.crisp-dm.org/).

Data modalities!

Structured Data

Un + Semi structured Data

Structured data

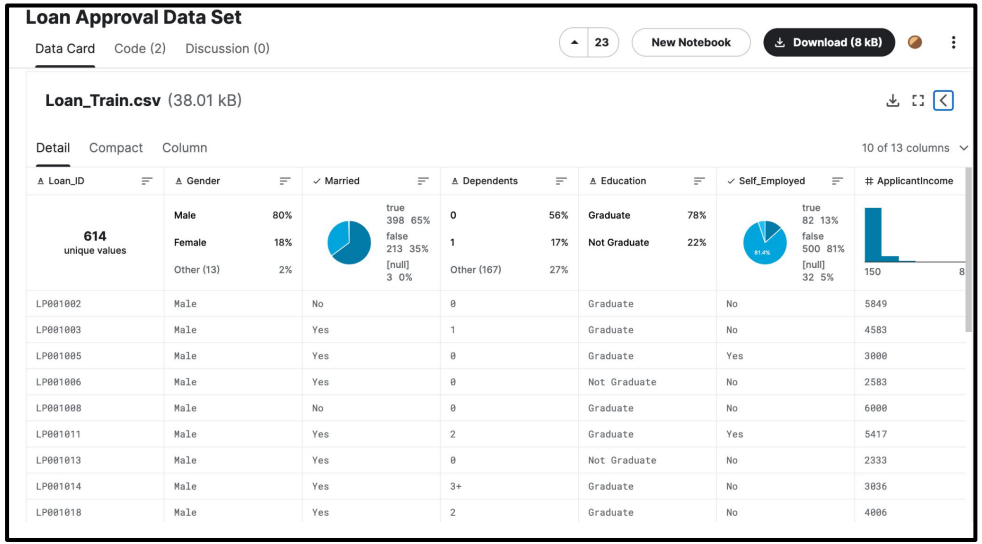


- Generally organized in **tables** and collected through filling **forms** (manual or online)
- Stored in **databases/spread-sheets** mainly
- Also popular the **.csv** file format

- Opening a bank account
- University registration
- Gmail
- Amazon account
- Your health profile

[Kaggle loan approval dataset]

Structured data



Loan Approval Data Set

Data Card | Code (2) | Discussion (0) | 23 | New Notebook | Download (8 kB)

Loan_Train.csv (38.01 kB)

Detail | Compact | Column | 10 of 13 columns

Loan_ID	Gender	Married	Dependents	Education	Self_Employed	ApplicantIncome
614 unique values	Male 80% Female 18% Other (13) 2%	true 398 65% false 213 35% [null] 3 0%	0 56% 1 17% Other (167) 27%	Graduate 78% Not Graduate 22%	true 82 13% false 500 81% [null] 32 5%	150 – 8
LP001002	Male	No	0	Graduate	No	5849
LP001003	Male	Yes	1	Graduate	No	4583
LP001005	Male	Yes	0	Graduate	Yes	3000
LP001006	Male	Yes	0	Not Graduate	No	2583
LP001008	Male	No	0	Graduate	No	6000
LP001011	Male	Yes	2	Graduate	Yes	5417
LP001013	Male	Yes	0	Not Graduate	No	2333
LP001014	Male	Yes	3+	Graduate	No	3036
LP001018	Male	Yes	2	Graduate	No	4006

- Generally collected through **forms** (manual or online)
- Stored in **databases/spread-sheets** mainly
- Also popular the **.csv** file format

- Opening a **bank account**
- **University registration**
- **Gmail, Azure, and/or Amazon** account
- Your immigration, health, social media **profile**

[Kaggle loan approval dataset]

Structured data

- There also other formats, or you can convert them to
- Some are stand alone, while others are sequences and/or series
- Software generated logs

- Genomics data
- Stock prices
- Your CC history
- Weather data
- Google maps

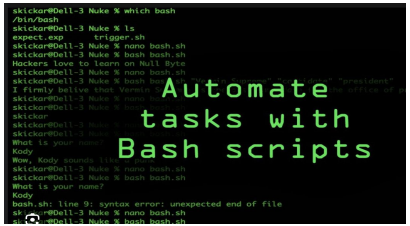
Structured data

- There also other formats, or you can convert them to
- Software generated logs
- Some are stand alone, white others are sequences and/or series

- Genomics data
- Stock prices
- Your CC history
- Weather data
- Google maps

Un/Semi Structured data

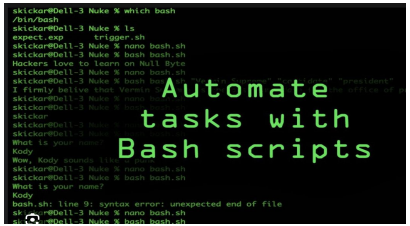
- Free forms
- Stored in data lake/warehouses mainly
- **Languages**: sequence of strings (semantics)
- **Audio**: Language + Acoustics; Music
- **Image** : Visual representation of the world
- **Video**: Sequence of images



- Some are sequence, while others are stand alone
- Social media data (emotions, vives)

Un/Semi Structured data

- Free forms
- Stored in data lake mainly?
- Languages: sequence of strings
- Audio: Language + Acoustics; Music
- Image : Visual representation of the world
- Video: Sequence of images



- Some are sequences, while others are stand alone
- Social media data (emotions, vives)

Un/Semi Structured data

- Free forms
- Stored in data lake mainly?
- Languages: sequence of strings
- Audio: Language + Acoustics; Music
- Image : Visual representation of the world
- Video: Sequence of images

- Some are sequence, while others are stand alone
- **Social media** data (discussions, messages, emotions, vives)

Python (DS focused)

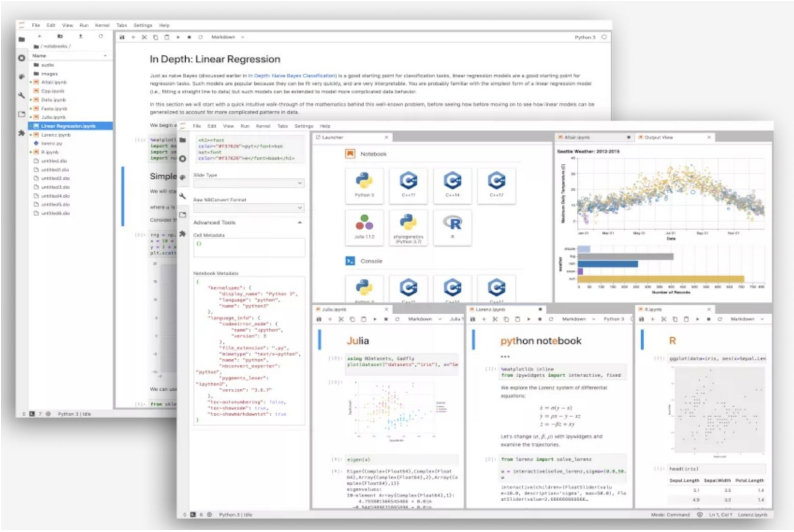
Data Processing

Visualization

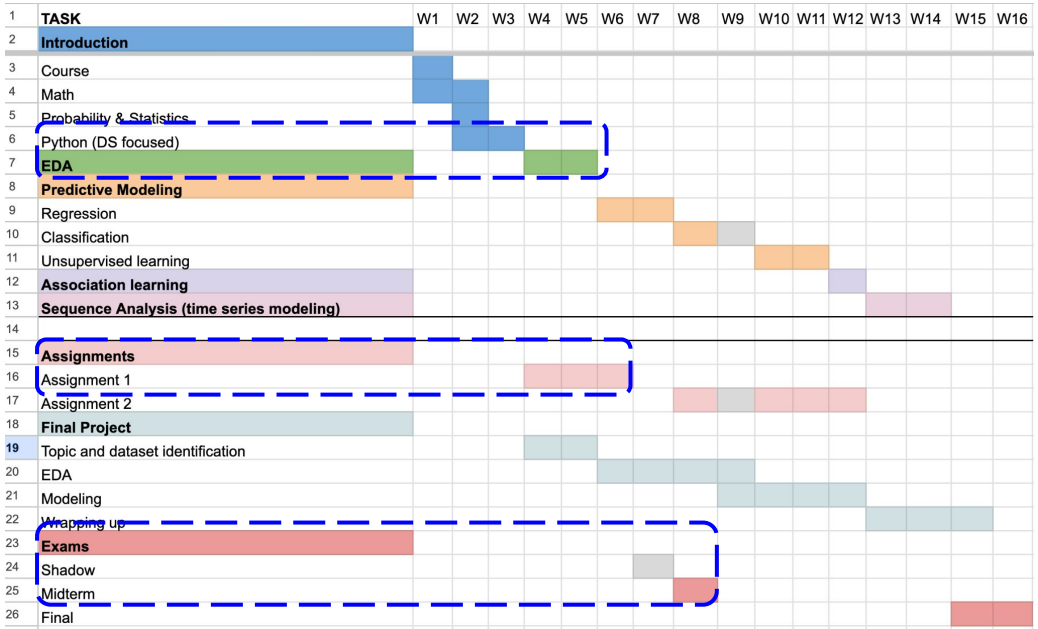
Prog. env: Google Colab (is a Jupyter Notebook)

- **Google** Colab(oratory) is a **Python** based Jupyter Notebook.
- **Jupyter Notebook** is an open source **client-server** based programming environment
  - Interactive programming experience (multi-language support) in the form of **documents**
  - Easy markdown and visualization capabilities
  - Cell based execution workflow. Documents contain both executable code and markdown texts/links/figures etc.
- **IPython** is the predecessor and **Jupyterlab** (next generation Notebook interface) is the successor.

Google Colaboratory is-a jupyter



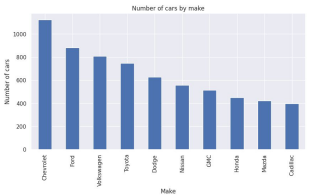
Exploratory Data Analysis (EDA)



Exploratory Data Analysis (EDA)

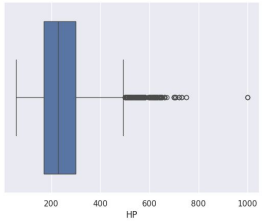
Automobile

- Data types
  - Object data type
- Null/missing value imputations
- Queries:
  - Unique values
  - Duplicate rows, columns
  - Estimating min, max, avg, mode

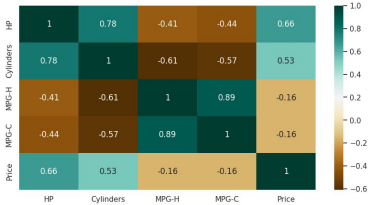


Frequency plot

Box plot



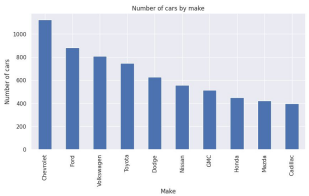
Correlation plot



Exploratory Data Analysis (EDA)

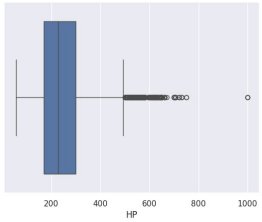
Automobile

- Data types
  - Object data type
- Null/missing value imputations
- Queries:
  - Unique values
  - Duplicate rows, columns
  - Estimating min, max, avg, mode

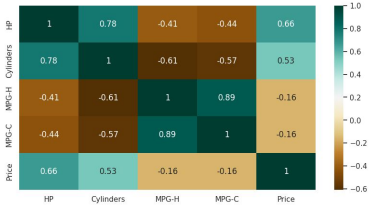


Frequency plot

Box plot



Correlation plot



Exploratory Data Analysis (EDA)

Retail

Indexing
- Records (rows, columns)

Date time data type

Data aggregations

How to bring information from multiple Tables/dataframes
- Joins (inner, left, right, outer)

joins

```
sales['Date'] = pd.to_datetime(sales.Date, format="%d/%m/%Y")
```

```
sales['Year'] = sales.Date.dt.year
sales['Month'] = sales.Date.dt.month
sales['Week'] = sales.Date.dt.isocalendar().week
```

Exploratory Data Analysis (EDA)

Retail

Indexing
- Records (rows, columns)

Date time data type

Data aggregations

How to bring information from multiple Tables/dataframes
- Joins (inner, left, right, outer)

joins

```
sales['Date'] = pd.to_datetime(sales.Date, format="%d/%m/%Y")
```

```
sales['Year'] = sales.Date.dt.year
sales['Month'] = sales.Date.dt.month
sales['Week'] = sales.Date.dt.isocalendar().week
```

Exploratory Data Analysis (EDA)

Retail

Indexing
- Records (rows, columns)

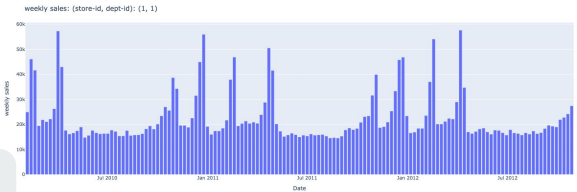
Date time data type

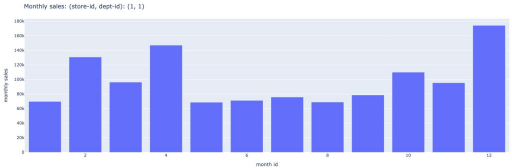
Data aggregations

How to bring information from multiple Tables/dataframes
- Joins (inner, left, right, outer)

joins

Weekly to monthly





Exploratory Data Analysis (EDA)

Retail

Indexing
- Records (rows, columns)

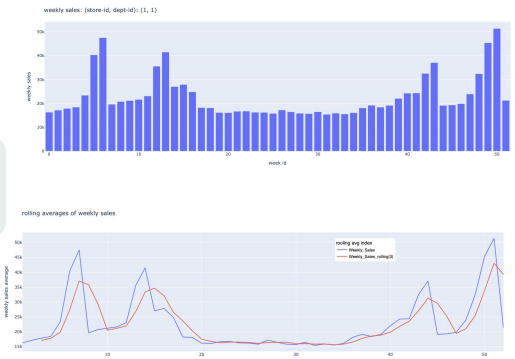
Date time data type

Data aggregations (smoothing)

How to bring information from multiple Tables/dataframes
- Joins (inner, left, right, outer)

joins

Weekly smoothing



Exploratory Data Analysis (EDA)

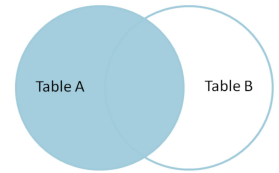
Indexing
- Records (rows, columns)

Date time data type

Data aggregations (smoothing)

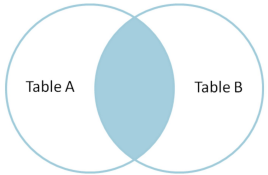
How to bring information from multiple Tables/dataframes
- **Joins (inner, left, right, outer)**

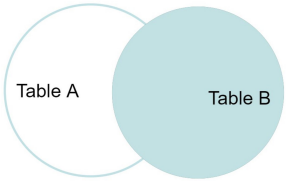
Table/Dataframe joins



left

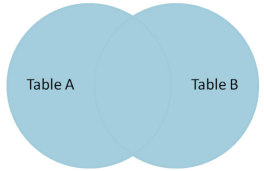
inner





right

outer



Assignment 1

1. [**3x2 points**] For each dimension $d \in [2^2, 2^4, ..., 2^{10}]$, sample 100 random points from corresponding vector spaces (sample code to generate random samples is provided below), and

   - Record the $l2$, $l1$, and the *cosine* distances between all pairs (of points); then
   - Fit three normal/Gaussian distributions, one for each distance metric. Also, share the mean (μ) and the standard distribution (σ) parameters of each distribution that you have learned.
   - Plot these normal/Gaussian distributions using your preferred visualization package(s).

   normal/Gaussian distribution: $p(x) \sim \mathcal{N}(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

   Sample code (to generate $n = 100$ random samples from a $d = 4$ dimensional vector space):

   ```
   import numpy as np
   d, n = 4, 100
   sample_data = np.random.randn(n, d)
   ```

2. [**3x1 points**] Probability :
   We provide you a set **male** and **female** student names, extracted from the NLTK corpus, in two different files below:

   female names file:
   https://raw.githubusercontent.com/mdkamrulhasan/data_mining_kdd/refs/heads/main/data/miscellaneous/female_nltk.csv
   male names file:
   https://raw.githubusercontent.com/mdkamrulhasan/data_mining_kdd/refs/heads/main/data/miscellaneous/male_nltk.csv

   We ask you to:

Assignment 1

(a) Estimate the **probability** of female-student names starting with "M".

(b) Estimate the **log-probability** of male-student names starting with "F".

(c) Estimate the **log-probability** of female or male-student names starting with "Z".

Note: For each of the above estimates, provide your results up to three (3) decimal points only.

3. [**11 points**] Climate Change Data Exploration:
Below, you are given links to some critical climate change related data in two different files.

climate distortion data file:
https://raw.githubusercontent.com/mdkamrulhasan/data_mining_kdd/refs/heads/main/data/climate-change/climate_distortion.csv

country-region data file:
https://raw.githubusercontent.com/mdkamrulhasan/data_mining_kdd/refs/heads/main/data/climate-change/country_region.csv

We ask you to answer following queries (each question is followed by it's corresponding score.)

(a) In how many columns, you detected NULL values in any of these data files? Impute those NULL values using your preferred method/logic? (1)

(b) How many unique Region(s) were recorded for USA? (1)

(c) Which country had the most and the least amount of "CO2_Emissions_MT" before and after 2000? What are these corresponding numbers? (2)

(d) What was the coolest date in the USA (based on the entire historical data recorded)? (1)

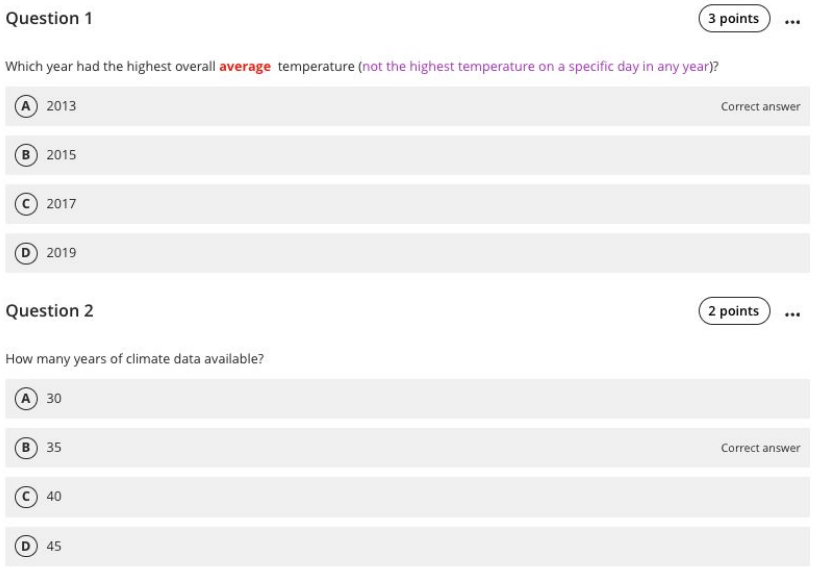
(e) What Crop_Type from the "Midwest(U)" and the "Northeast(U)" Regions together in the USA had the highest historical economic impact ("Economic_Impact_Million_USD")? (2)

(f) What was the yearly average precipitation ('Total_Precipitation_mm') in USA? Display these yearly average precipitation values through a graph. (2)

(g) Which of the following variable pairs has the highest and the lowest associations? If you find a tie, pick one from those. (2)

- Average_Temperature_C,
- Total_Precipitation_mm,
- CO2_Emissions_MT,
- Extreme_Weather_Events,
- Pesticide_Use_KG_per_HA,
- Fertilizer_Use_KG_per_HA,
- Soil_Health_Index,
- Economic_Impact_Million_USD

Midterm (EDA)


Question 1
3 points
Which year had the highest overall average temperature (not the highest temperature on a specific day in any year)?
A 2013
Correct answer
B 2015
C 2017
D 2019
Question 2
2 points
How many years of climate data available?
A 30
B 35
Correct answer
C 40
D 45

Midterm (EDA)

Question 3 — 2 points

Which country had the least number of **Extreme_Weather_Events**?

- (A) USA
- (B) Russia — Correct answer
- (C) China
- (D) India

Question 4 — 3 points

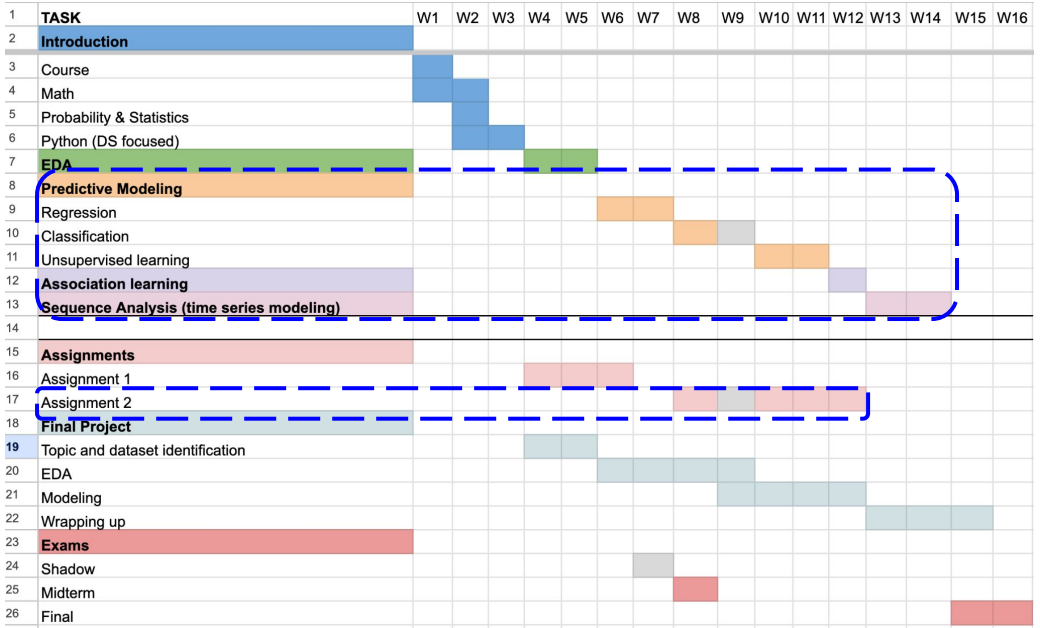
Which **Indian** region had the highest irrigation access area in percentage (Irrigation_Access_%)?

- (A) Maharashtra — Correct answer
- (B) West Bengal
- (C) Punjab
- (D) Tamil Nadu

Predictive modeling

Predictive modeling

Assignment 2



	TASK	W1	W2	W3	W4	W5	W6	W7	W8	W9	W10	W11	W12	W13	W14	W15	W16
2	**Introduction**																
3	Course																
4	Math																
5	Probability & Statistics																
6	Python (DS focused)																
7	**EDA**																
8	**Predictive Modeling**																
9	Regression																
10	Classification																
11	Unsupervised learning																
12	**Association learning**																
13	**Sequence Analysis (time series modeling)**																
14																	
15	**Assignments**																
16	Assignment 1																
17	Assignment 2																
18	**Final Project**																
19	Topic and dataset identification																
20	EDA																
21	Modeling																
22	Wrapping up																
23	**Exams**																
24	Shadow																
25	Midterm																
26	Final																

From 678 (compressed + aligned)

General ML models

Supervised

Regression

- kNN
- Linear Regression
- Decision Tree
- Meta learners
  - Random Forest Regressor
  - Boosting Regressor
- Support Vector Regressor (SVRs)

Classification

- kNN
- Logistic Regression
- Decision Tree
- Meta Learners
  - Random Forest Classifier
  - Boosting Classifiers
- Support Vector Classifiers (SVCs)
- Naive Bayes

General ML models

Unsupervised

Clustering

- Clustering algorithms
  - **k-means**: Centroid Based
  - k-modes: Mode Based (categorical)
  - **Hierarchical clustering**: Distance connectivity based
  - **GMM**: Distribution based
  - **DBSCAN**: Density Based
- How to choose the optimal number of clusters.

Linear Dimensionality Reduction

- Principal Component Analysis (PCA)
- Singular Value Decomposition (SVD)

General ML models

Model generalization

- Universal concepts (applies to all models)
  - Cross validation
  - HP optimization

Universal concepts

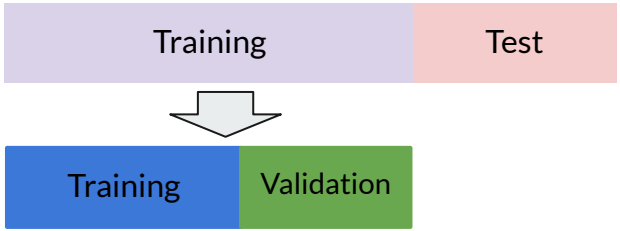
- Overfitting
- Underfitting

Overfitting vs Under fitting

General ML models

Model generalization

- Training set, Validation set, Test set
- Iid data

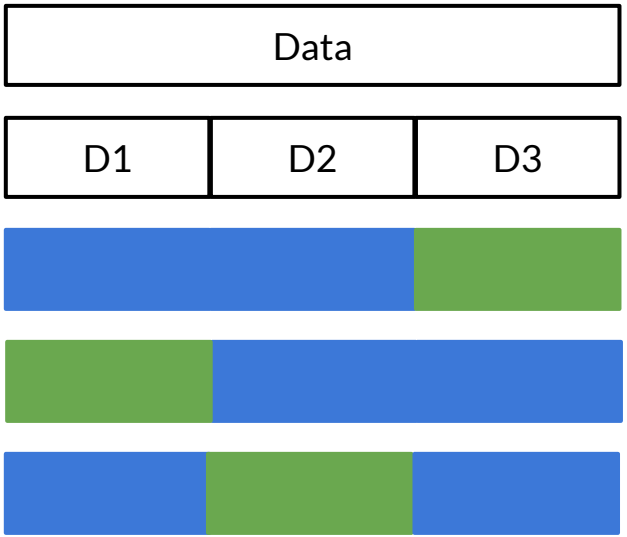


Data splits

- Overfitting
- Underfitting

Overfitting vs Under fitting

K-fold-cross validation



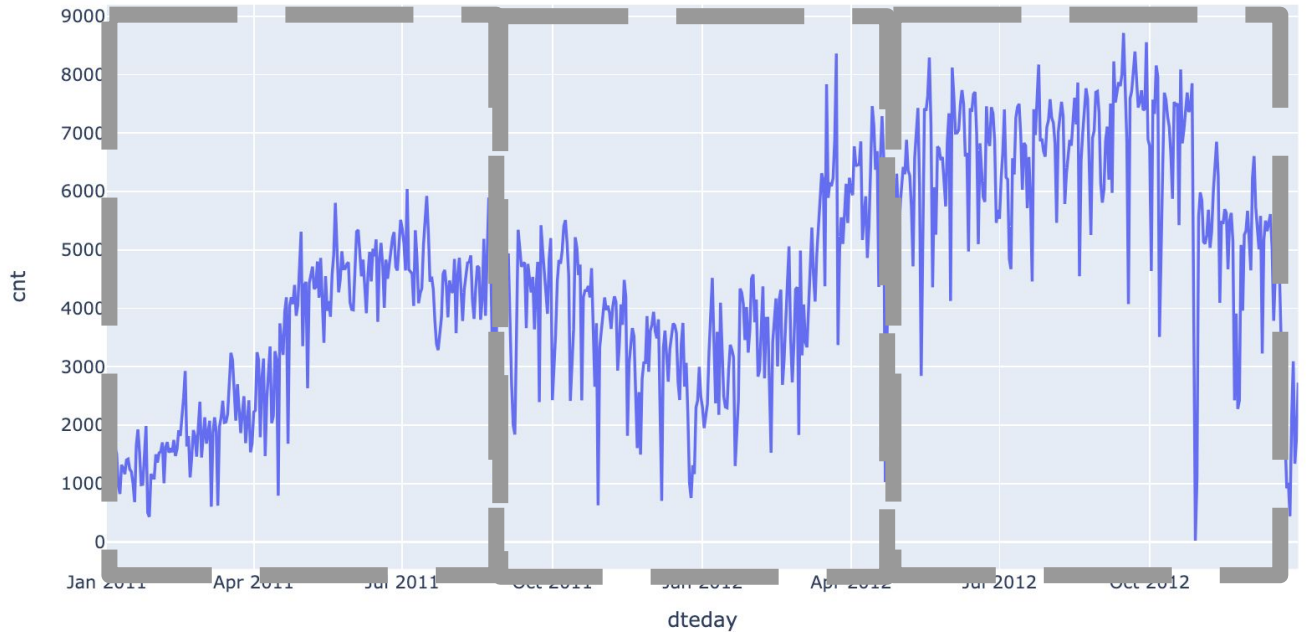
3-fold-cv

Train

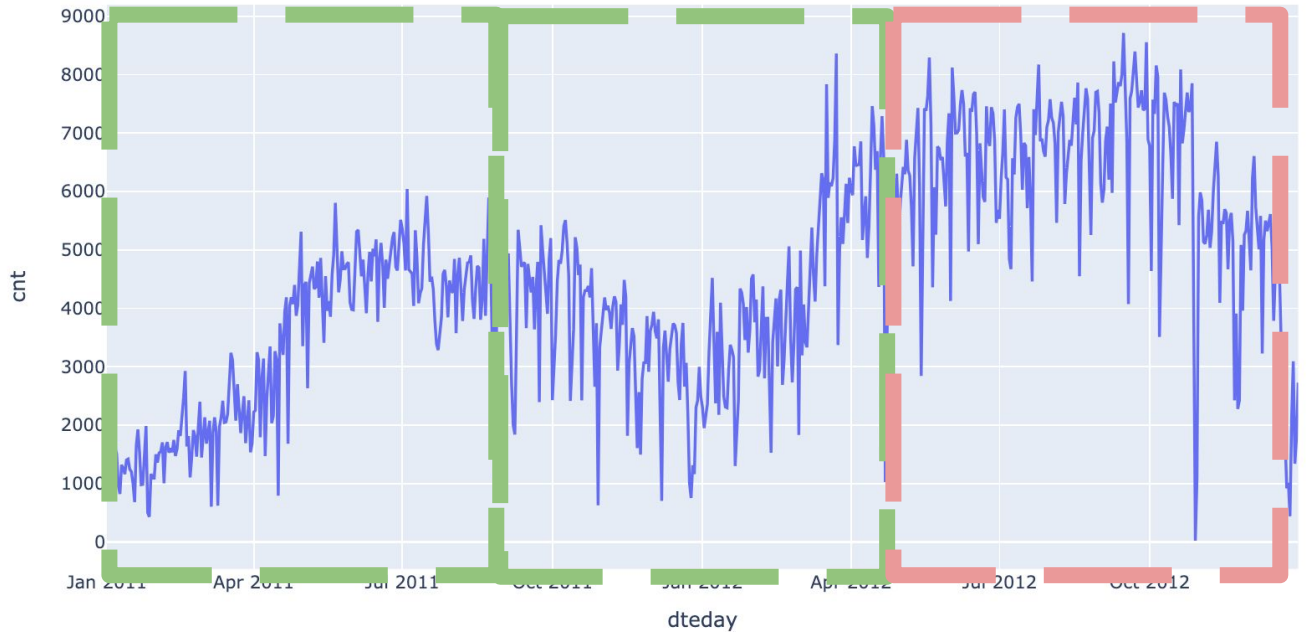
validate

What HP gives the best validation score?

Sequential cross validation



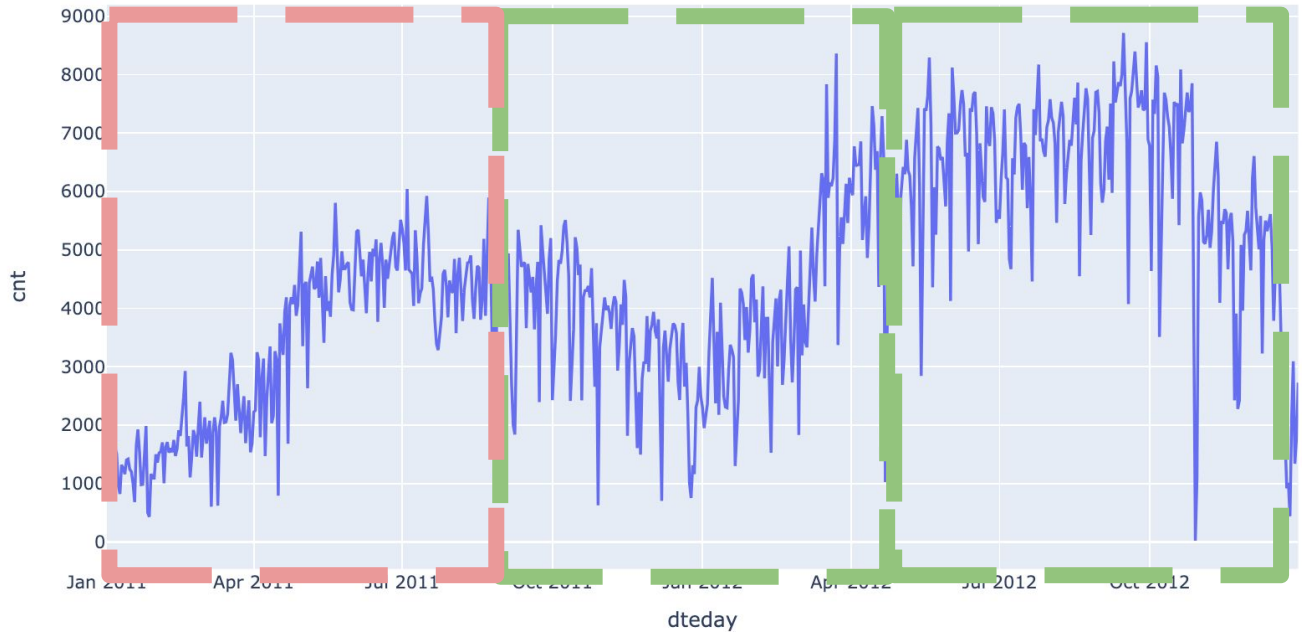
Sequential cross validation



Valid configuration:

*Training on **first two folds**, and validate on the **last fold***

Sequential cross validation

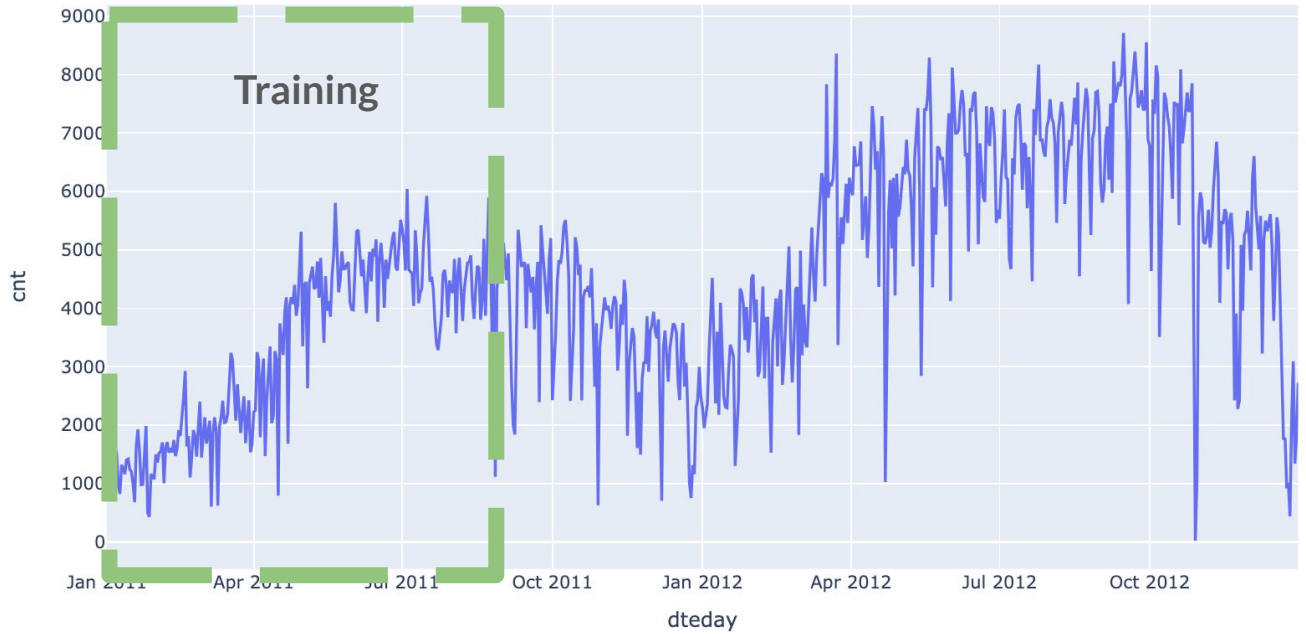


Invalid configuration:

*Training on **first fold**, and validate on the last two.*

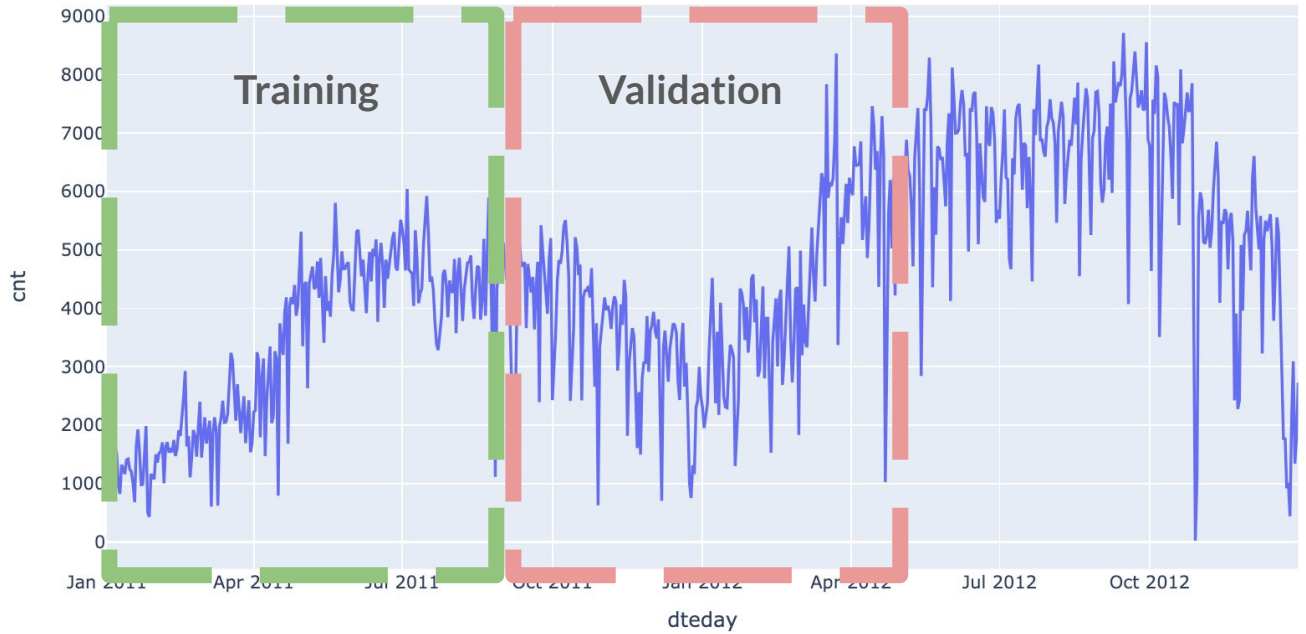
Sequential cross validation

Sequential cross validation



Always follow!

Sequential cross validation



Always follow!

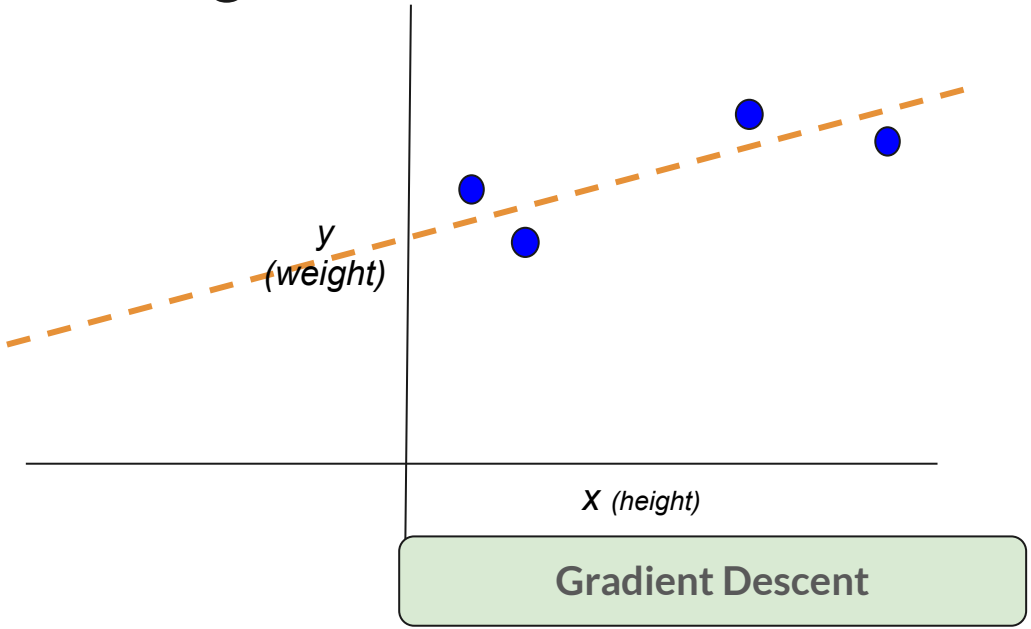
Sequential cross validation

Always follow!

Sequential cross validation

Always follow!

Linear Regression



Gradient Descent

Model

$$\hat{y} = \beta_0 + \beta_1 x$$

$$\Theta = \{\beta_0, \beta_1\}$$

Fitting Error

$$\epsilon = |\hat{y} - y|$$

Optimization function

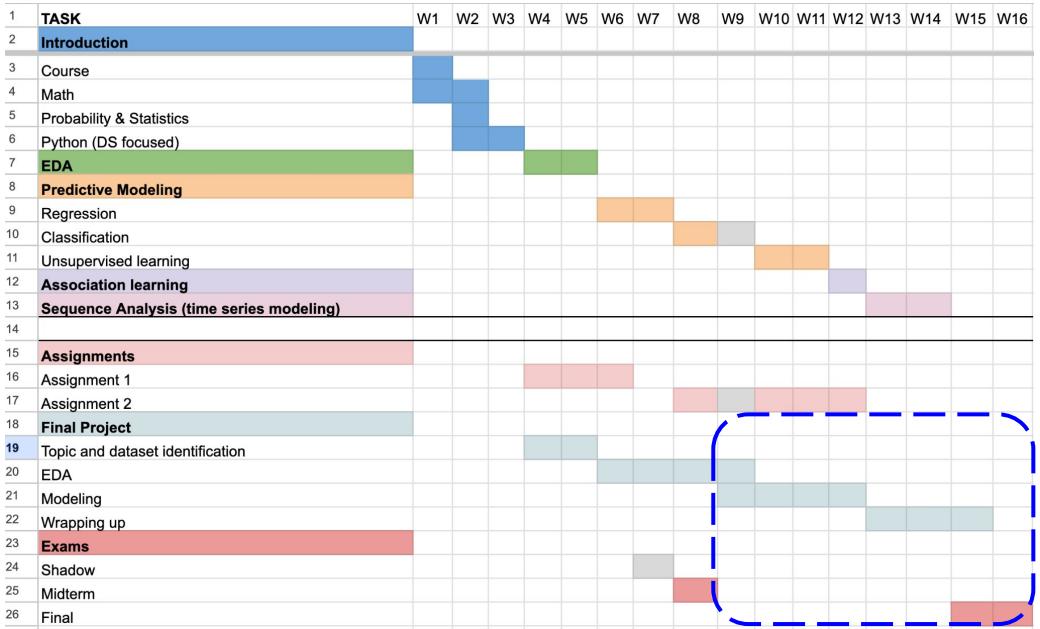
$$E_\Theta = \frac{1}{2}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2$$

$$\Theta^* = \text{argmin}_\Theta \, E\{(x_i, y_i)\}_{i=1,\cdots,N}$$

Course Wrap up + Final Exams

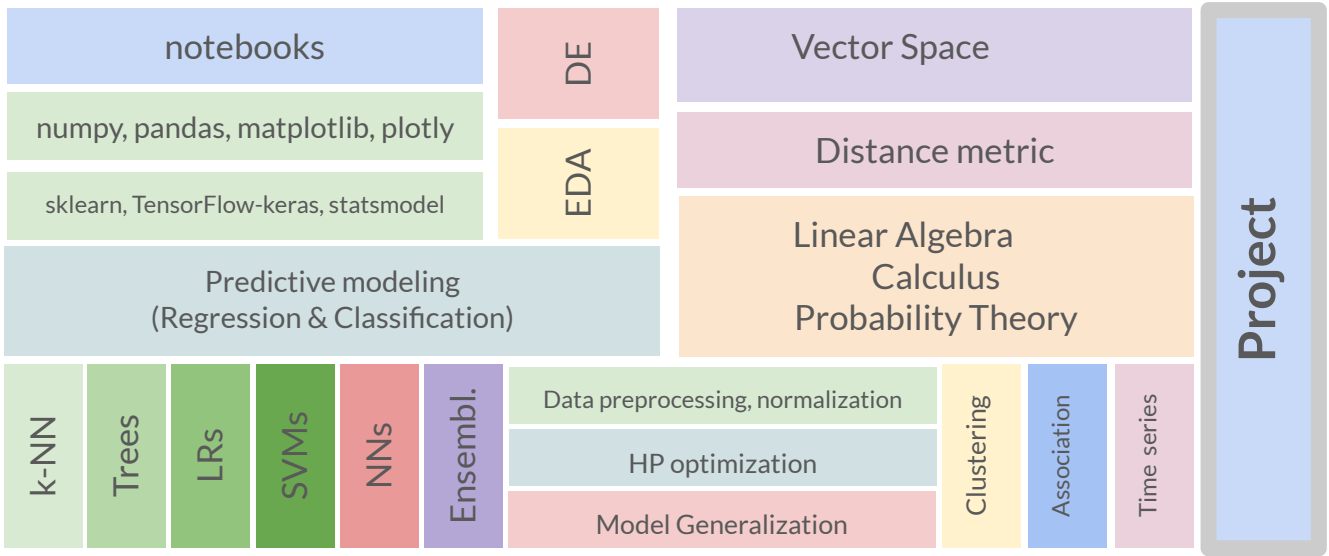
Course wrap-up

Final exam



QA

Key concepts!



Data modalities!

- Tabular data
-

Structured Data

- Language data
-

Un + Semi structured Data

Exploratory Data Analysis (EDA)

Retail:
Indexing
- Records (rows, columns)

Date time data type
Data aggregations
How to bring information from multiple Tables/dataframes
- Joins (inner, left, right, outer)

joins