zenius × Kampus Merdeka INDONESIA JAYA

Getting Started to Data Science

22 April 2022

Data Warehousing, Analysis, and Visualization for Business Insights

Program Studi Independen Bersertifikat
Zenius Bersama Kampus Merdeka

Quick Intro

Rahadian Rizki Prayoga

Education:

- Sekolah Tinggi Ilmu Statistik, Major : Stats, Minor : Economics

Roles:

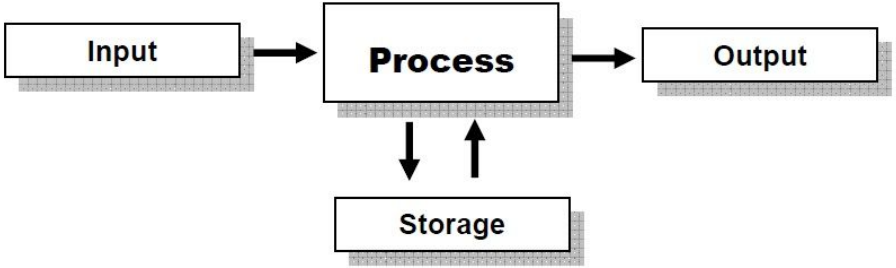
- **Data Analytics Lead - Enterprise Wholesale Div., Telkom Indonesia**
- Lead Data Scientist, Mamikos
- Vice Lead Big Data Analytics, Sinarmas Bank Tbk
- Senior Data Scientist, Akseleran
- Guest Lecturer - AI Subject, Universitas Gadjah Mada

https://www.linkedin.com/in/rahadianrizki/

1. Important Concepts
2. Roles in Data Science
3. How to Build Your Portfolio
4. Intro to Kaggle and Github
5. Other Best Practices

Important Concepts

Data Analytics

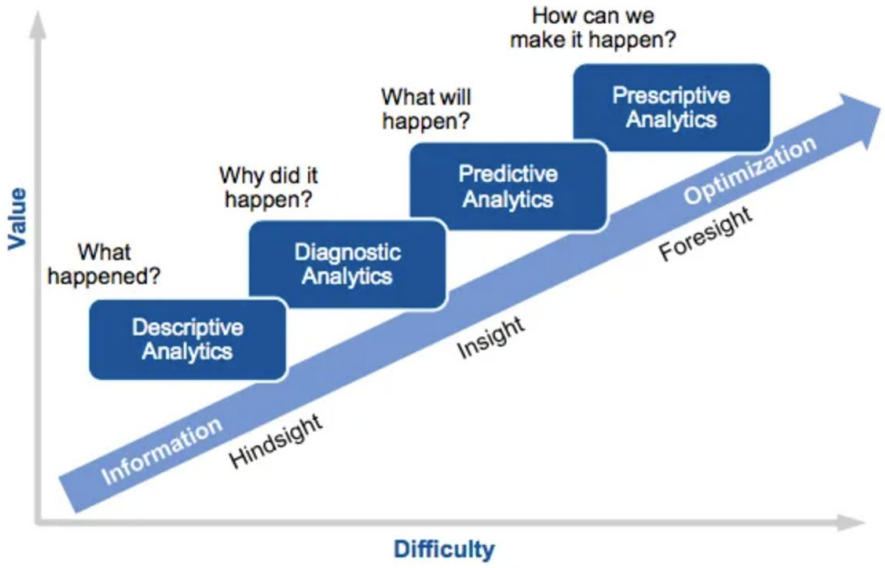


Data Analytics

"Garbage in, garbage out"

Your analysis is as good as your data.

Data Analytics



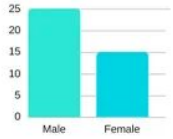
Source : Gartner Analytics Ascendancy Model
https://www.clickz.com/how-can-ai-allow-marketers-to-predict-the-future/112268/gartner-analytic-ascendancy-model/
https://www.gartner.com/en/topics/data-and-analytics

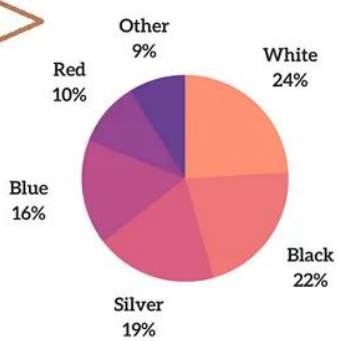
Descriptive Analytics

- Describing the data
- Common Calculation :
  - Sums
  - Counts
  - Averages
- Typical Reports :
  - Tables
  - Bar Charts
  - Pie Charts
  - Narratives

	A	B	C	D
1	Respondent Number	Age	Gender	Favorite Car Color
2	1	22	M	White
3	2	37	F	Silver
4	3	45	F	Black
5	4	62	F	Gray
6	5	28	M	Red
7	6	45	M	Green
8	7	88	F	Brown
9	8	61	M	White
10	9	95	M	Black
11	10	27	M	White
12	11	39	F	Green
13	12	43	M	Brown
14	13	55	F	Black
15	14	59	F	White

RAW DATA

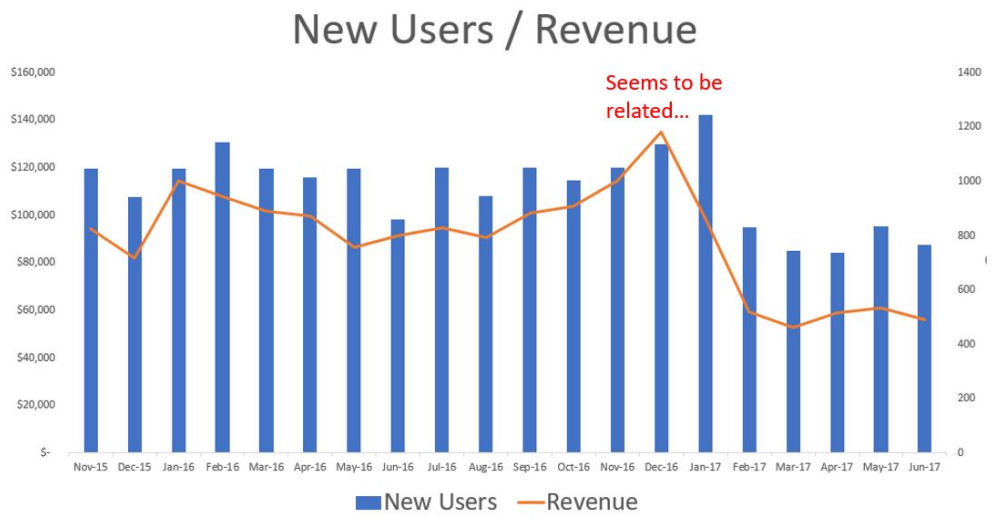




Descriptive Statistics

Diagnostic Analytics

- Answers “Why did it happen?”
- Drill Down Techniques
- Data Discovery
- Correlations
- Combining Charts
- Discover Related Metrics



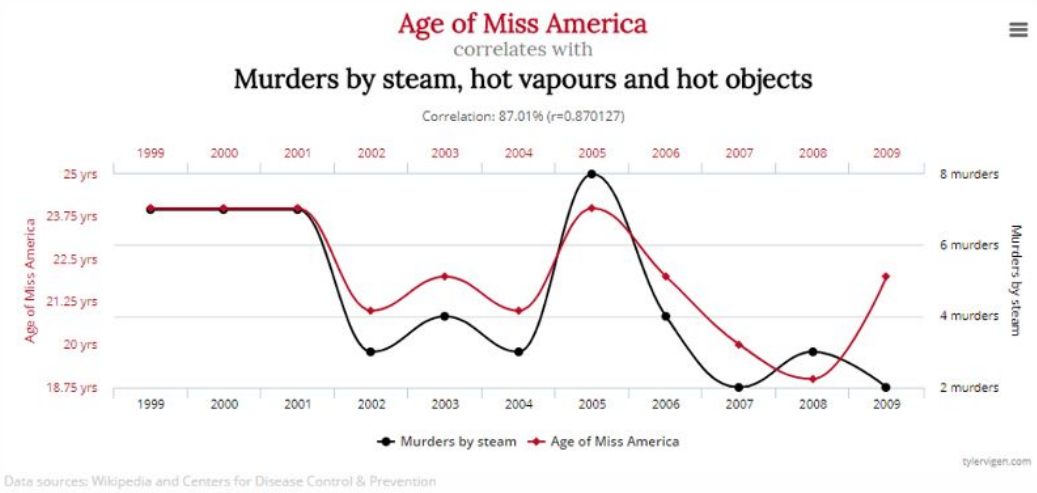
Diagnostic Analytics

Correlation **doesn't prove** Causation

Correlation will tell you when two variables (say clicks and conversions) move **in sync** with one another

While it's tempting to draw conclusions from that fact, the **correlation must** also **make sense** before it can be considered as **causal evidence**

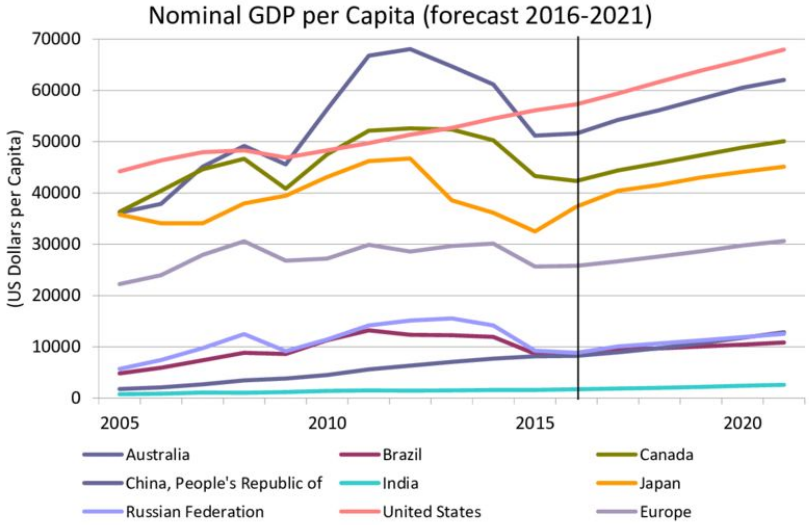
That's why we need **Business Acumen**.



Predictive Analytics

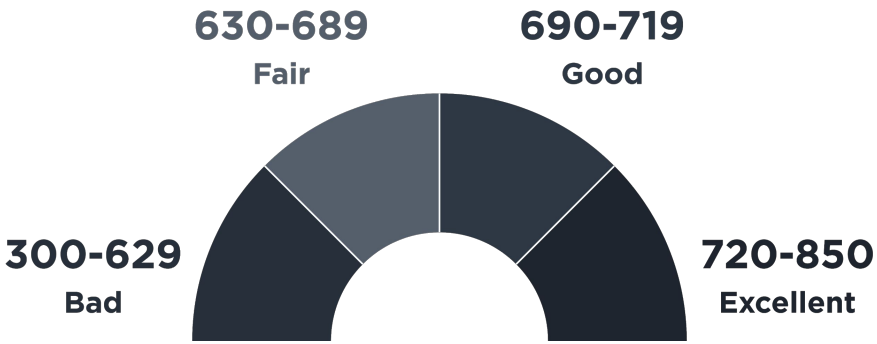
a variety of **statistical techniques** from **data mining**, **predictive modelling**, and **machine learning**, that analyze **current** and h**istorical facts** to make **predictions about future** or otherwise unknown events.

- **exploiting patterns** found in historical and transactional data
- **identifying** risks and opportunities
- **capturing relationships** among many factors to the target
- **guiding** decision-making



Predictive Analytics

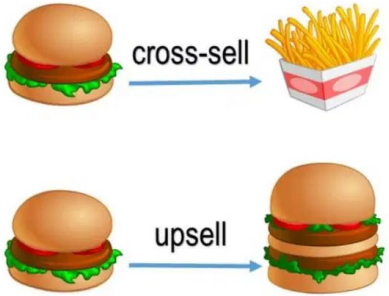
Credit Risk Scoring



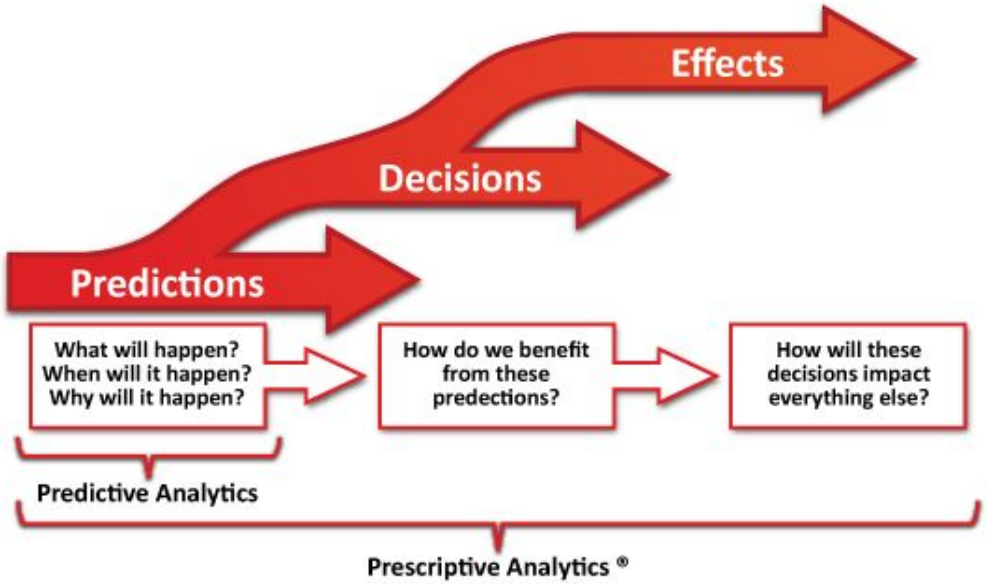
Sentiment Analysis

Word	Sentiment
good	0.5
great	0.8
terrible	-0.8
alright	0.1

Cross-Selling



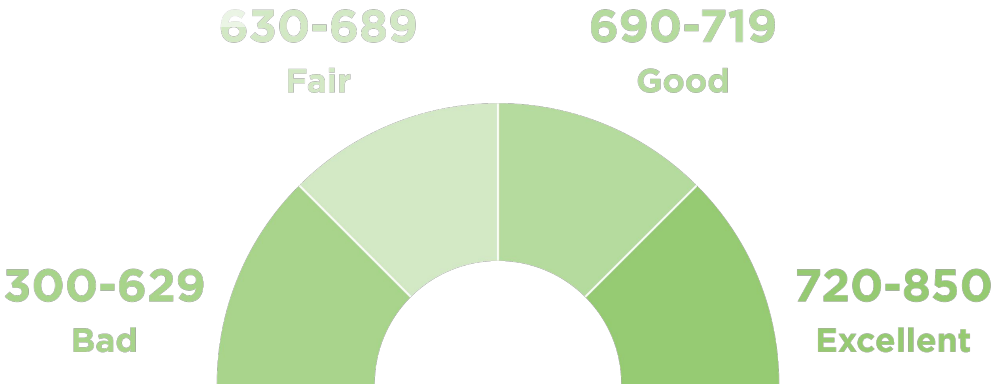
Prescriptive Analytics



also include **Optimization.**

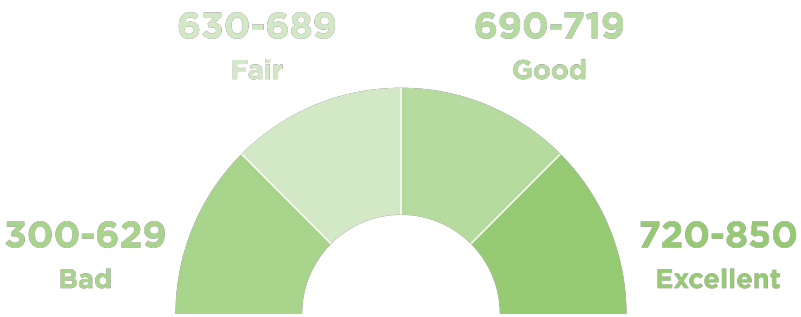
Prescriptive Analytics

Credit Risk Scoring



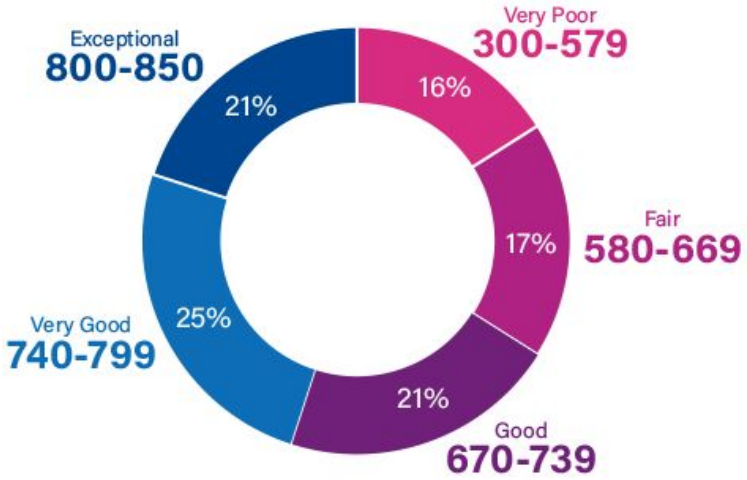
- How much is the **Expected Credit Loss (ECL)** ?
- How about the **Probability of Default (PD)** ?
- Where is **the best cut-off** for Bad and Good given X risk appetite ?
- When someone is **accepted** for a loan, will someone with **840** credit score has the same **LTV** as other one who has **700** ?

Prescriptive Analytics



Customer with no credit history

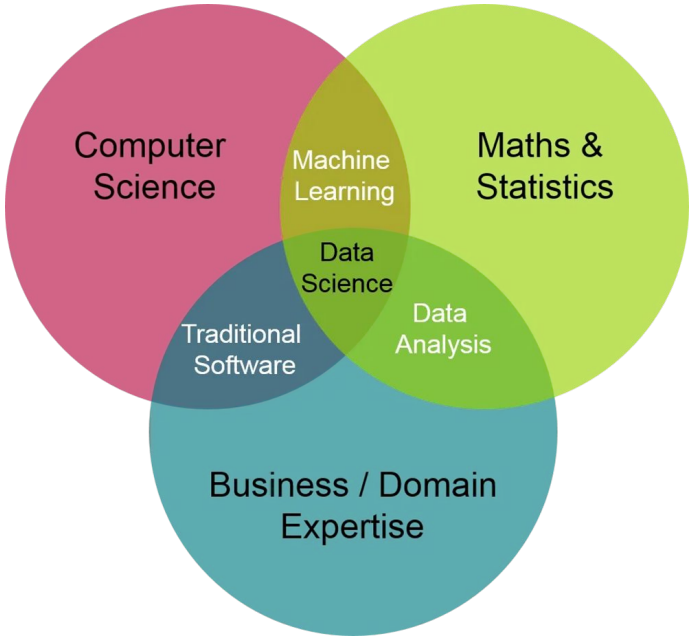
Customer with credit history



Different Product, Different Credit Scorecard
Different Region, Different Credit Scorecard
Unbankable vs Bankable Customer Credit Scorecard

Roles in Data Science

Roles in Data Science

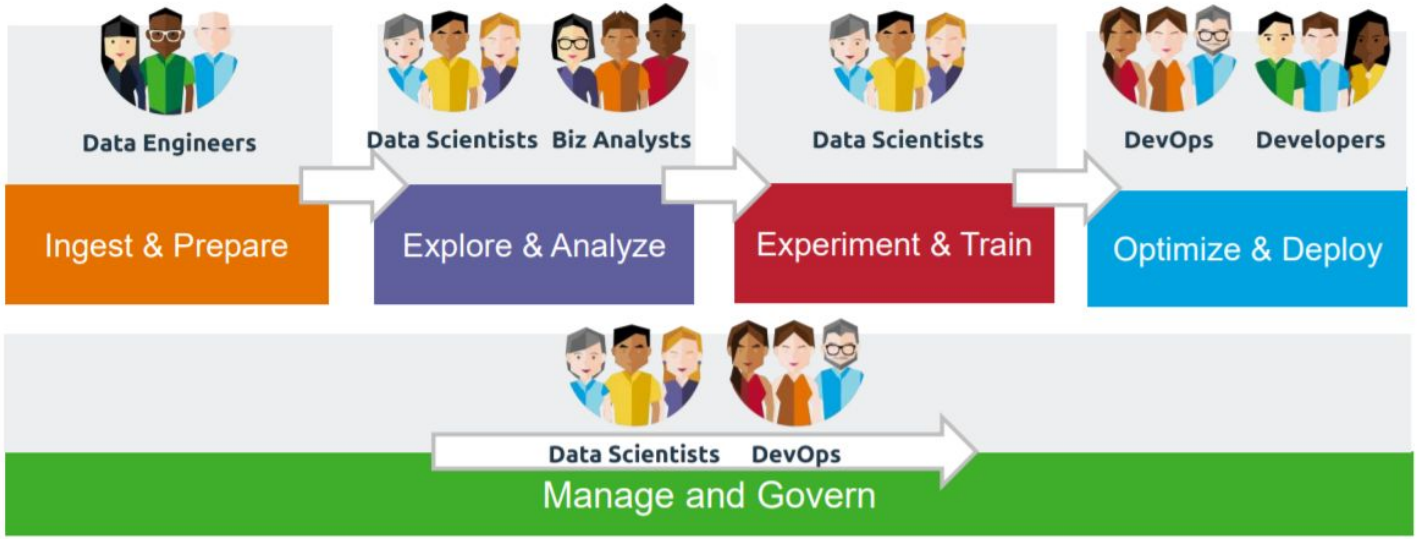


Roles in Data Science

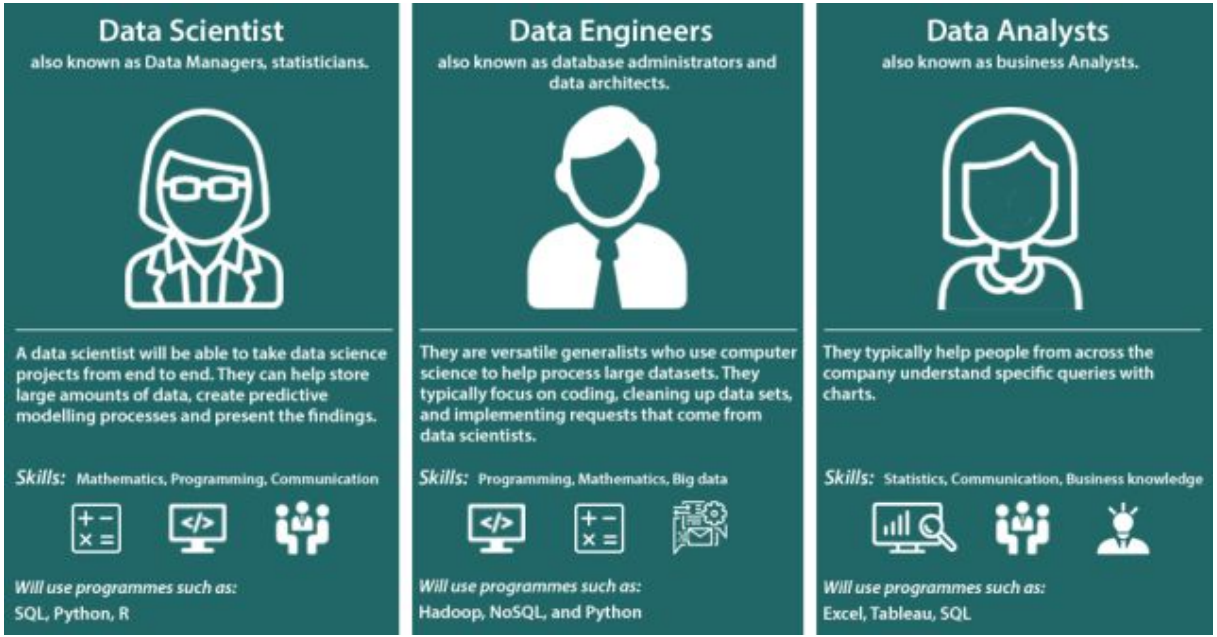
"A data scientist is someone who is better at statistics than any programmer and better at programming than any statistician"

- a random stranger on twitter

Data Science : Team Sport



Data Scientist



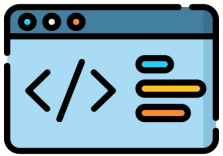
Data Analyst

Analytics

Problem Solving, Data Exploration

Visualization

Right Plot for The Right Purpose



Programming & Tools

SQL, Python, Excel

Statistics

Uni/Bi/Multi-variate, Hypothesis Testing

Business Acumen

Understanding The Subject Matter Deeply

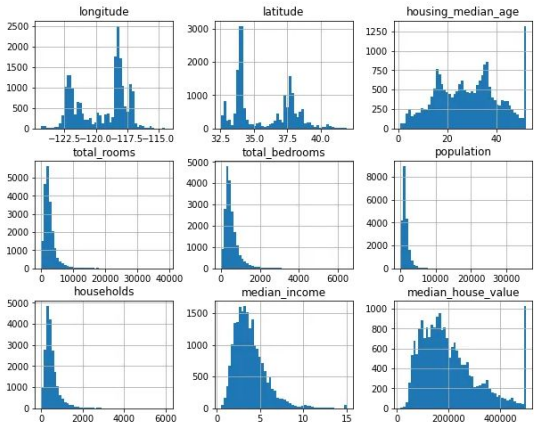
How to Build Your Portfolio

How To Build Your Portfolio

Building a portfolio is an essential part to conquer the career struggle.

- **Choose a topic you're interested in**
  Don't get complicated, make sure the topic to analyze is something you know well, to ease the way.
- **End-to-end**
  Make a complete portfolio from start to finish
- **Explainable**
  Make sure your audience/interviewer able to understand what you're making and how you make it.
- **Make a story and publish it!**
  Use platform like kaggle, github, medium and linkedin to spread the awesome stuff and the journey!

Example of a Good Portfolio and Projects



https://amankharwal.medium.com/130-machine-learning-projects-solved-and-explained-605d188fb392

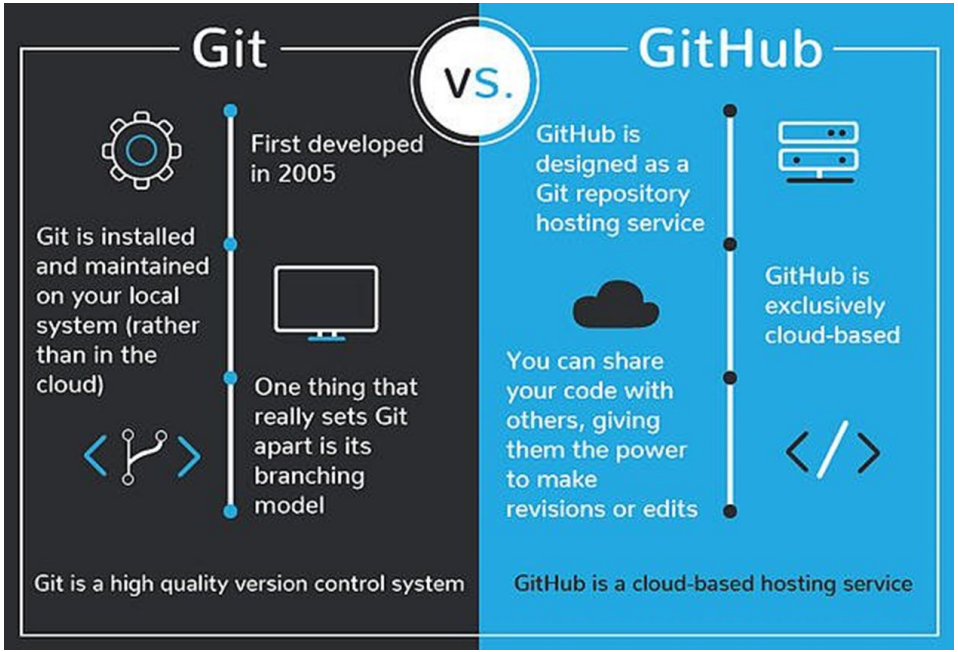
***credit to the owner.**

Intro to Kaggle and Github

Github

At a high level, GitHub is a website and cloud-based service that helps developers store and manage their code, as well as track and control changes to their code.

Github



Version Control

Version Control is the general term.

Version control lets developers safely work through branching and merging.

With **branching**, a developer duplicates part of the source code (called the repository). The developer can then safely make changes to that part of the code without affecting the rest of the project.

Then, once the developer gets his or her part of the code working properly, he or she can **merge (merging)** that code back into the main source code to make it official.

All of these changes are then tracked and can be reverted if need be.

Git

Git is a specific open-source version control system created by Linus Torvalds in 2005.

Specifically, **Git** is a distributed version control system, which means that the entire codebase and history is available on every developer's computer, which allows for easy branching and merging.

Let's try to create a github account and do some demos there ! Link : https://github.com/

Complete Github Tutorial for Portfolio :

https://chriskhanhtran.github.io/_posts/2020-01-13-portfolio-tutorial/

Github

command	description
`git clone` ***`url [dir]`***	copy a git repository so you can add to it
`git add` ***`files`***	adds file contents to the staging area
`git commit`	records a snapshot of the staging area
`git status`	view the status of your files in the working directory and staging area
`git diff`	shows diff of what is staged and what is modified but unstaged
`git help` ***`[command]`***	get help info about a particular command
`git pull`	fetch from a remote repo and try to merge into the current branch
`git push`	push your new branches and data to a remote repository
others: `init`, `reset`, `branch`, `checkout`, `merge`, `log`, `tag`	

Some of git commands.

Let's try it and do some demo!

Kaggle

Kaggle offers a no-setup, customizable, Jupyter Notebooks environment. Access free GPUs and a huge repository of community published data & code.

https://www.kaggle.com/

Kaggle

Predict Malicious Websites: XGBoost *Draft saved*

File Edit Insert Run View Help

+ Add dataset | Commit

```python
data = pd.read_csv("../input/dataset.csv")

# clean up column names
data.columns = data.columns.\
    str.strip().\
    str.lower()

# remove non-numeric columns
data = data.select_dtypes(['number'])

# split data into training & testing
train, test = train_test_split(data, shuffle=True)

# peek @ dataframe
train.head()
```

+ Code + Markdown

```python
# split training data into inputs & outputs
X = train.drop(["type"], axis=1)
Y = train["type"]

# specify model (xgboost defaults are generally fine)
model = xgb.XGBRegressor(tree_method = "gpu_exact")

# fit our model
model.fit(y=Y, X=X)
```

In[]:

```python
# split testing data into inputs & output
test_X = test.drop(["type"], axis=1)
test_Y = test["type"]

# predictions & actual values, from test set
predictions = model.predict(test_X) > 0
actual = test_Y
```

Sessions

Draft Session | 1m/9h | GPU Off

CPU	0.01%
RAM	237.3MB/16GB
Disk	266.3MB/4.9GB

Workspace

input (read-only data)

Versions

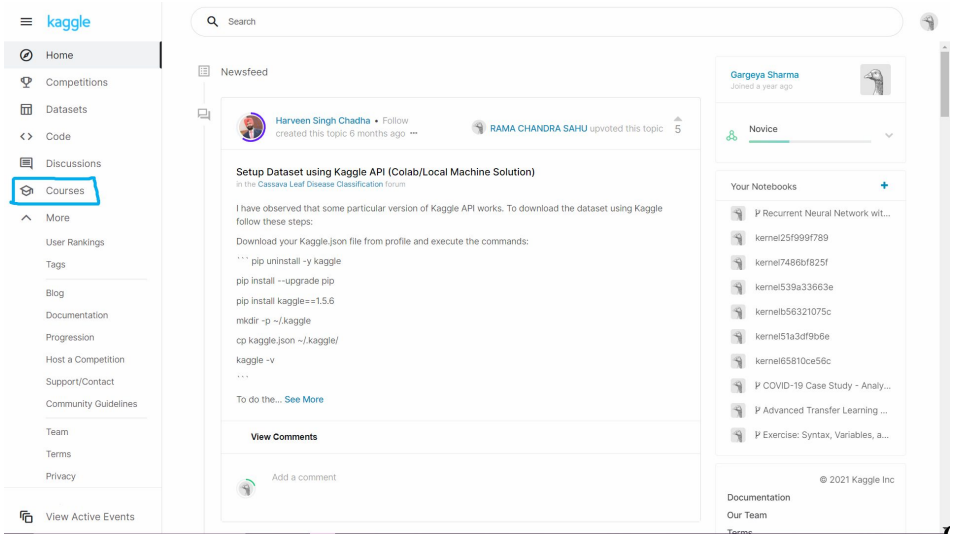
Settings

Sharing	Private
Language	Python
Docker	Latest Available
GPU	On
Internet	Off
Packages	Install...
BigQuery	Enable...

Console Draft Session (0m10s) | CPU 45% | GPU Off | RAM 4.5/8GB | Disk 32MB/128GB

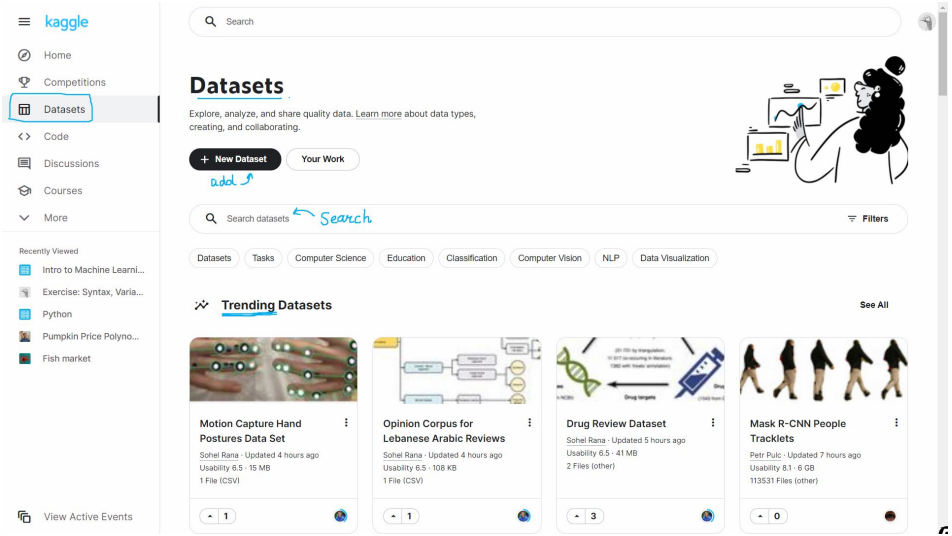
Advantages of Using Kaggle

1. **Free Courses and free certificates available**



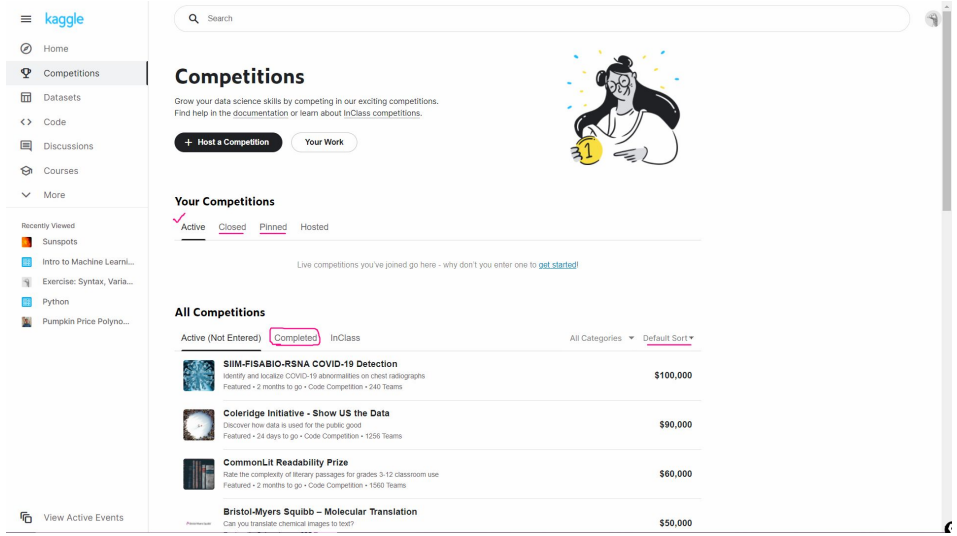
Advantages of Using Kaggle

2. A Huge collection of publicly available/ contributed datasets to practice/ work on



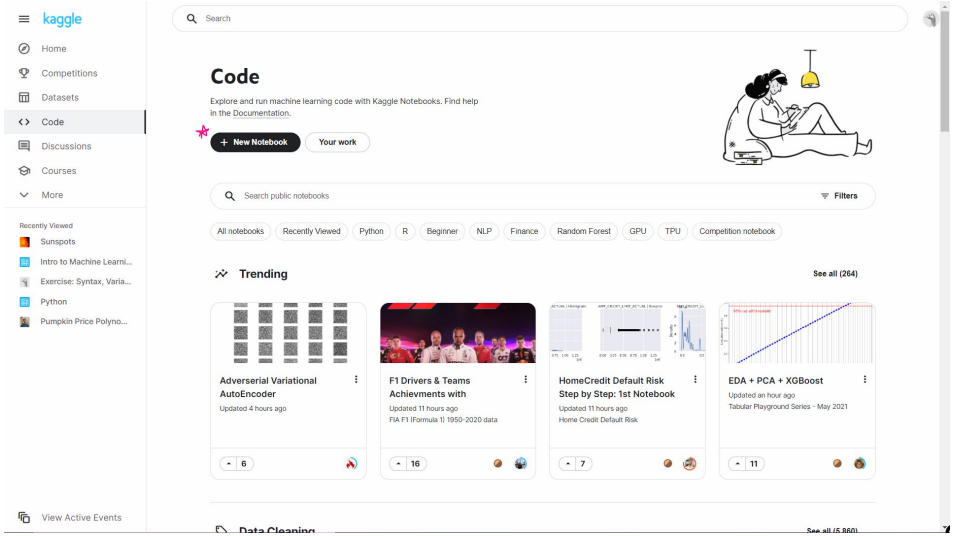
Advantages of Using Kaggle

3. Data Science/ Machine Learning / Deep learning Competitions



Advantages of Using Kaggle

4. Kaggle Notebooks / Code



Other Best Practices

Let’s talk about how to google!

- **Use Quotes To Match Exact Phrases**
  “pandas groupby”
- **Use AND/OR Operators**
  keyword1 and keyword2, **ex** : hadoop and hivesql, python or r
- **Exclude Certain Terms Using the Minus Sign**
  ex : ruby -gemstone
- **Use Wild Cards in Your Search Term**
  phrase

  ex : how to use * in pandas
- **Find Websites Similar to Another Website**
  related:[website_url], **ex** : related:kaggle.com
- **Search a Website Using Google**
  keyword1 site:[website_url] , **ex** : site:udemy.com machine learning
- **Find Content in a Specific File Type**
  keyword1 filetype:[file type], ex : filetype:pdf naive bayes
- **Use “before” and “after” Operators**
  keyword1 before:[date] **, ex** : pandas groupby after:2021

https://betterprogramming.pub/11-tricks-to-master-the-art-of-googling-as-a-software-developer-2e00b7568b7d

Let’s try to do the demo on our beloved search engine!

Building Portfolio and Resume ?

1. **Length** - Do keep it simple and one page max
2. **Objective** - Don't include one
3. **Coursework** - Do list relevant coursework
4. **Skills** - Don't give numerical ratings to your skills
5. **Skills** - Do list technical skills that job mentions
6. **Projects** - Don't list common projects or homework
7. **Projects** - Do show results and include links
8. **Portfolio** - Do fill out your online presence
9. **Experience** - Do tailor your experience towards the job

https://www.youtube.com/watch?v=xrhPjE7wHas **(Kaggle Channel)**

Thanks!
Any Questions?