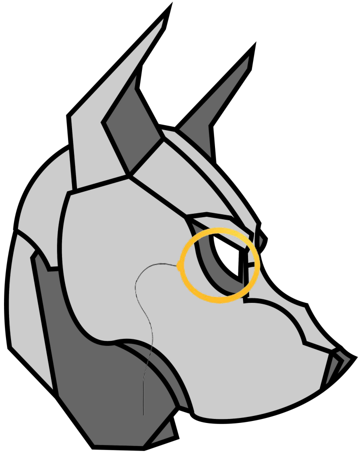
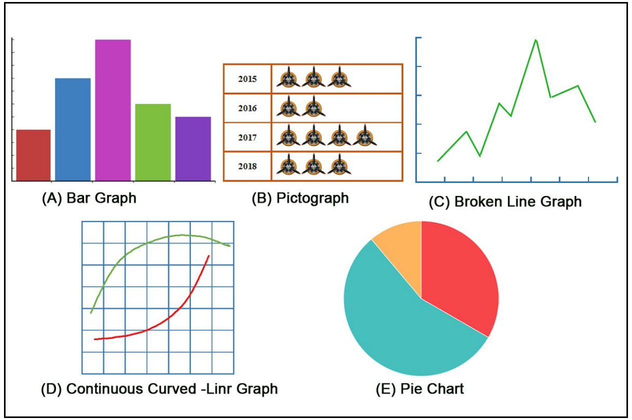
Cyber Defense Organization

Fall 2021 - Practical Data Analysis





DISCLAIMER

I do not claim to be an expert, these are the insights from what I have learned through coursework, my internship, and outside studying.

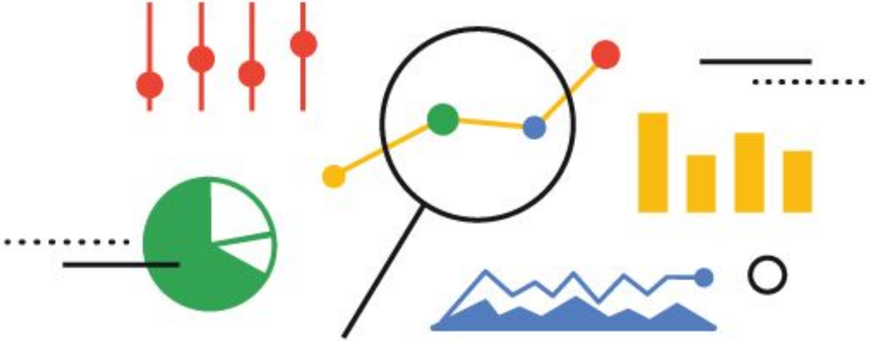
From my experience, I have found this to be an *extremely* valuable soft-skill.

Keep in mind, there are entire full time positions in many companies purely dedicated to data science and analysis.

Why data analysis?

Decisions, Decisions...

- The ultimate goal of any good data analysis is to gather conclusions to gain insight or to affect a decision.
- Relatively un-mentioned in cybersecurity and related fields.
- Can be a super valuable skill to have at your disposal.
- You'll undoubtedly have the best presentations.



This can be simple and complex

This could be as simple from using basic Excel functions and using pivot tables, to analyzing text sentiment with nltk in python.

Data

vs

Information

vs

Knowledge

Data:

“Raw, unanalyzed, unorganized, unrelated, uninterrupted material generated from which is used to derive information after analysis”

Essentially plain facts, observations, or statistics.

Information:

“Data that has been processed, aggregated, or organized”

Gives significance and meaning to data points

Knowledge:

“the state of knowing something with cognizance through the understanding of concepts, study and experience”

How we apply information in a use to take action, develop insights, and make decisions

The End Goal

“See the forest through the trees”

Using data

To gain information

To make decisions

Where Does Data Analysis Come In?

Basically, with data analysis we are trying to make sure we can make informed decisions.

Data analytics is almost more of a comp-sci field, but there is some overlap to security.

How does this apply to cyber security?

Application of Data Analysis

Prediction using machine learning

Automation!

Decision making

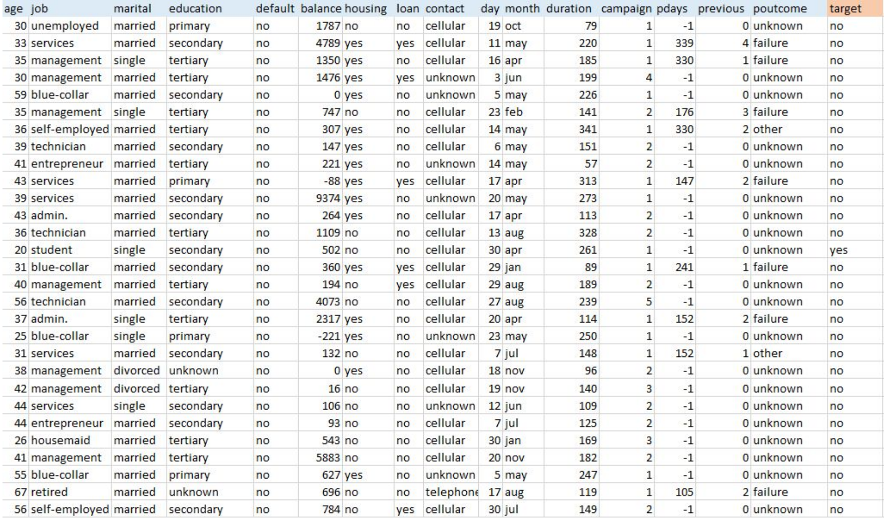
A lot of *things* generate data

Be it logs, or any other tool that makes a lot of data.

Like, lots of data.

It could be 20 rows or 20,000,000.

250,000 rows of data is useless if you can’t gather any conclusions from it.



age	job	marital	education	default	balance	housing	loan	contact	day	month	duration	campaign	pdays	previous	poutcome	target
30	unemployed	married	primary	no	1787	no	no	cellular	19	oct	79	1	-1	0	unknown	no
33	services	married	secondary	no	4789	yes	yes	cellular	11	may	220	1	339	4	failure	no
35	management	single	tertiary	no	1350	yes	no	cellular	16	apr	185	1	330	1	failure	no
30	management	married	tertiary	no	1476	yes	yes	unknown	3	jun	199	4	-1	0	unknown	no
59	blue-collar	married	secondary	no	0	yes	no	unknown	5	may	226	1	-1	0	unknown	no
35	management	single	tertiary	no	747	no	no	cellular	23	feb	141	2	176	3	failure	no
36	self-employed	married	tertiary	no	307	yes	no	cellular	14	may	341	1	330	2	other	no
39	technician	married	secondary	no	147	yes	no	cellular	6	may	151	2	-1	0	unknown	no
41	entrepreneur	married	tertiary	no	221	yes	no	unknown	14	may	57	2	-1	0	unknown	no
43	services	married	primary	no	-88	yes	yes	cellular	17	apr	313	1	147	2	failure	no
39	services	married	secondary	no	9374	yes	no	unknown	20	may	273	1	-1	0	unknown	no
43	admin.	married	secondary	no	264	yes	no	cellular	17	apr	113	2	-1	0	unknown	no
36	technician	married	tertiary	no	1109	no	no	cellular	13	aug	328	2	-1	0	unknown	no
20	student	single	secondary	no	502	no	no	cellular	30	apr	261	1	-1	0	unknown	yes
31	blue-collar	married	secondary	no	360	yes	yes	cellular	29	jan	89	1	241	1	failure	no
40	management	married	tertiary	no	194	no	yes	cellular	29	aug	189	2	-1	0	unknown	no
56	technician	married	secondary	no	4073	no	no	cellular	27	aug	239	5	-1	0	unknown	no
37	admin.	single	tertiary	no	2317	yes	no	cellular	20	apr	114	1	152	2	failure	no
25	blue-collar	single	primary	no	-221	yes	no	unknown	23	may	250	1	-1	0	unknown	no
31	services	married	secondary	no	132	no	no	cellular	7	jul	148	1	152	1	other	no
38	management	divorced	unknown	no	0	yes	no	cellular	18	nov	96	2	-1	0	unknown	no
42	management	divorced	tertiary	no	16	no	no	cellular	19	nov	140	3	-1	0	unknown	no
44	services	single	secondary	no	106	no	no	unknown	12	jun	109	2	-1	0	unknown	no
44	entrepreneur	married	secondary	no	93	no	no	cellular	7	jul	125	2	-1	0	unknown	no
26	housemaid	married	tertiary	no	543	no	no	cellular	30	jan	169	3	-1	0	unknown	no
41	management	married	tertiary	no	5883	no	no	cellular	20	nov	182	2	-1	0	unknown	no
55	blue-collar	married	primary	no	627	yes	no	unknown	5	may	247	1	-1	0	unknown	no
67	retired	married	unknown	no	696	no	no	telephone	17	aug	119	1	105	2	failure	no
56	self-employed	married	secondary	no	784	no	yes	cellular	30	jul	149	2	-1	0	unknown	no

As you spend time in security…

There are countless tools that generate a lot of ***data***, not all tools generate ***information*** and some generate poor information.

Some tools just generate a ridiculous amount of data.

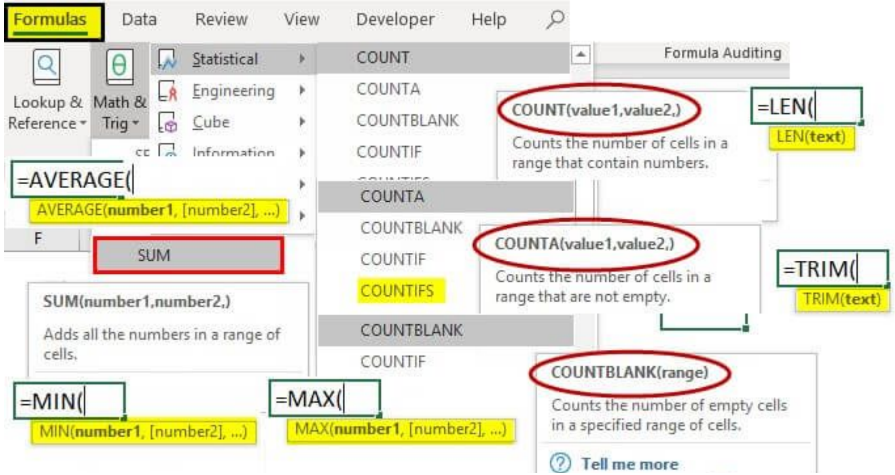
Simple

Excel is probably the most basic we can get.

However, once we start pushing Excel with larger data sets, it can get slow and not perform well.

Excel does not often lend itself to more complex operations such as text analysis. Do not discount Excel!

Basic Excel Formulas



COUNT(value1,value2,)

Counts the number of cells in a range that contain numbers.

COUNTA(value1,value2,)

Counts the number of cells in a range that are not empty.

SUM(number1,number2,)

Adds all the numbers in a range of cells.

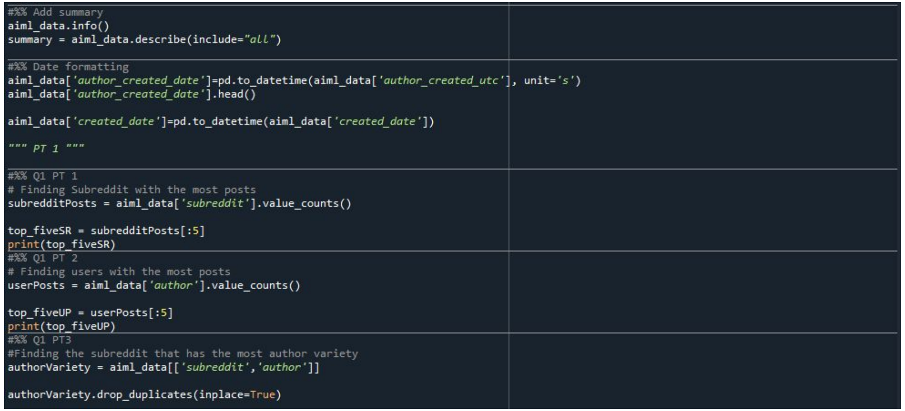
COUNTBLANK(range)

Counts the number of empty cells in a specified range of cells.

More Complex

For more basic functions and smaller data sets, excel will almost always be faster.

Python and Pandas come in when we wish to implement some more advanced analysis and functionality.



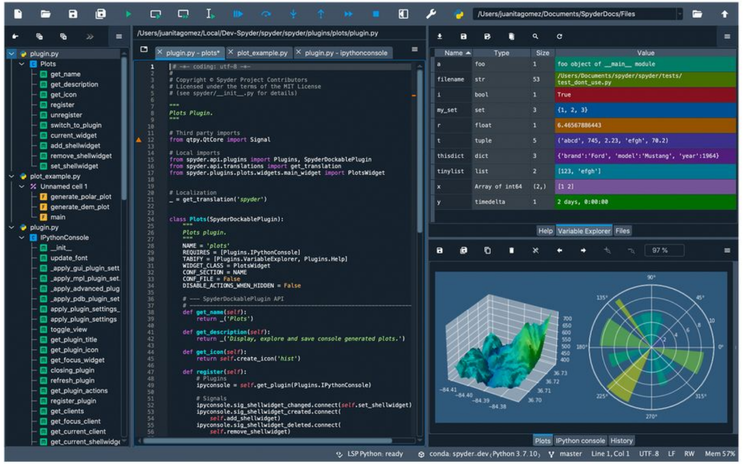
```
#%% Add summary
aiml_data.info()
summary = aiml_data.describe(include="all")

#%% Date formatting
aiml_data['author_created_date']=pd.to_datetime(aiml_data['author_created_utc'], unit='s')
aiml_data['author_created_date'].head()

aiml_data['created_date']=pd.to_datetime(aiml_data['created_date'])

""" PT 1 """

#%% Q1 PT 1
# Finding Subreddit with the most posts
subredditPosts = aiml_data['subreddit'].value_counts()

top_fiveSR = subredditPosts[:5]
print(top_fiveSR)
#%% Q1 PT 2
# Finding users with the most posts
userPosts = aiml_data['author'].value_counts()

top_fiveUP = userPosts[:5]
print(top_fiveUP)
#%% Q1 PT3
#Finding the subreddit that has the most author variety
authorVariety = aiml_data[['subreddit','author']]

authorVariety.drop_duplicates(inplace=True)
```

IDE of Choice

I like spyder, there is nothing wrong with other IDEs but I find the features that spyder offers to be rather comprehensive and it is essentially purpose built with analysis in mind.

spyder

The Scientific Python Development Environment



/Users/juanitagomez/Local/Dev-Spyder/spyder/spyder/plugins/plots/plugin.py

```python
# -*- coding: utf-8 -*-
#
# Copyright © Spyder Project Contributors
# Licensed under the terms of the MIT License
# (see spyder/__init__.py for details)

"""
Plots Plugin.
"""

# Third party imports
from qtpy.QtCore import Signal

# Local imports
from spyder.api.plugins import Plugins, SpyderDockablePlugin
from spyder.api.translations import get_translation
from spyder.plugins.plots.widgets.main_widget import PlotsWidget


# Localization
_ = get_translation('spyder')


class Plots(SpyderDockablePlugin):
    """
    Plots plugin.
    """
    NAME = 'plots'
    REQUIRES = [Plugins.IPythonConsole]
    TABIFY = [Plugins.VariableExplorer, Plugins.Help]
    WIDGET_CLASS = PlotsWidget
    CONF_SECTION = NAME
    CONF_FILE = False
    DISABLE_ACTIONS_WHEN_HIDDEN = False

    # --- SpyderDockablePlugin API
    # ------------------------------------------------------------------------
    def get_name(self):
        return _('Plots')

    def get_description(self):
        return _('Display, explore and save console generated plots.')

    def get_icon(self):
        return self.create_icon('hist')

    def register(self):
        # Plugins
        ipyconsole = self.get_plugin(Plugins.IPythonConsole)

        # Signals
        ipyconsole.sig_shellwidget_changed.connect(self.set_shellwidget)
        ipyconsole.sig_shellwidget_created.connect(
            self.add_shellwidget)
        ipyconsole.sig_shellwidget_deleted.connect(
            self.remove_shellwidget)
```

/Users/juanitagomez/Documents/SpyderDocs/Files

Name	Type	Size	Value
a	foo	1	foo object of __main__ module
filename	str	53	/Users/Documents/spyder/spyder/tests/test_dont_use.py
i	bool	1	True
my_set	set	3	{1, 2, 3}
r	float	1	6.46567886443
t	tuple	5	('abcd', 745, 2.23, 'efgh', 70.2)
thisdict	dict	3	{'brand':'Ford', 'model':'Mustang', 'year':1964}
tinylist	list	2	[123, 'efgh']
x	Array of int64	(2,)	[1 2]
y	timedelta	1	2 days, 0:00:00

LSP Python: ready conda: spyder-dev (Python 3.7.10) master Line 1, Col 1 UTF-8 LF RW Mem 57%

Spyder Features

Editor

Work efficiently in a multi-language editor with a function/class browser, code analysis tools, automatic code completion, horizontal/vertical splitting, and go-to-definition.

IPython Console

Harness the power of as many IPython consoles as you like in one GUI. Run code by line, cell or file; or work interactively with debugging, plots and magic commands.

Variable Explorer

Interact with and modify variables on the fly: plot a histogram or timeseries, edit a dateframe or Numpy array, sort a collection, dig into nested objects, and more!

Plots

Browse, zoom, copy and save the figures and images you create.

Debugger

Trace each step of your code's execution interactively.

Help

Instantly view any object's docs, and render your own.

Benefits of Python & Pandas

- Built to handle large data sets, good for ambitious projects.
- You can use other Python libraries in combination with Pandas. NumPy, SciPy, MatPlotLib and more!
- Pandas has so many features built in. Honestly, a rather ridiculous amount.
- Good data representation makes it easier to understand what you are doing.

Downsides

- Steep learning curve (kind of) - Counter point that the documentation is pretty good and StackOverflow is your friend.
- If you wanted to do 3D matrices, then you want to look into NumPy instead.
- You have to know Python, or at least be willing to learn.

Learning and Direction

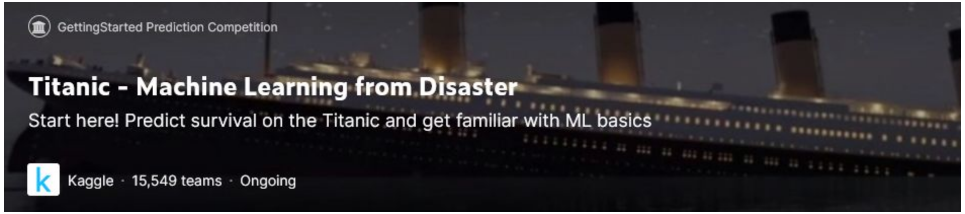
Projects

There are plenty of resources that provide data science projects and data sets for you to learn. Also, money.

https://www.kaggle.com/

Kaggle is a great resource and they even have competitions which are wonderful learning opportunities.

Here are some of the ones going on at the moment:

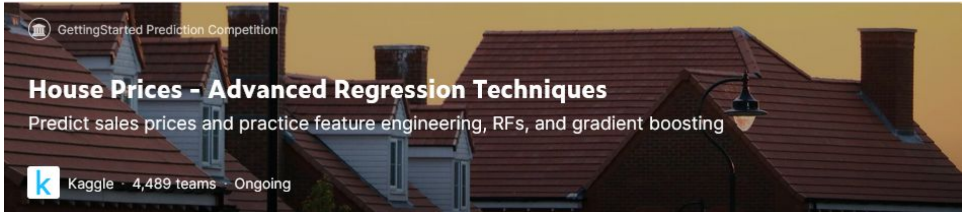


GettingStarted Prediction Competition

Titanic - Machine Learning from Disaster

Start here! Predict survival on the Titanic and get familiar with ML basics

Kaggle · 15,549 teams · Ongoing

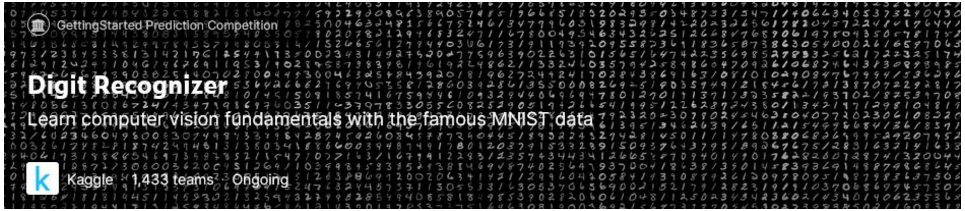


GettingStarted Prediction Competition

House Prices - Advanced Regression Techniques

Predict sales prices and practice feature engineering, RFs, and gradient boosting

Kaggle · 4,489 teams · Ongoing



GettingStarted Prediction Competition

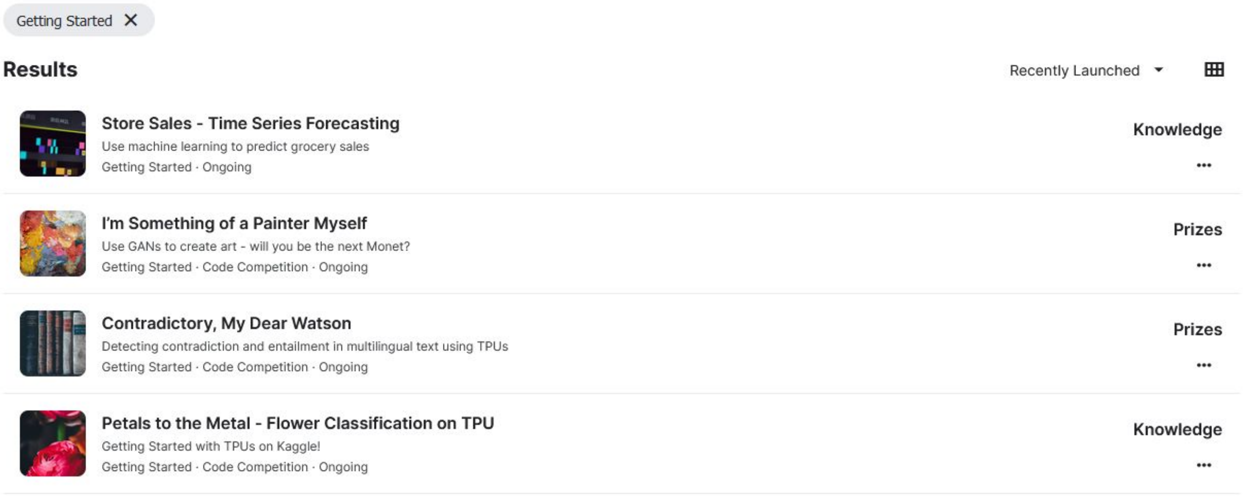
Digit Recognizer

Learn computer vision fundamentals with the famous MNIST data

Kaggle · 1,433 teams · Ongoing

Projects cont.

There are so many projects, there is bound to be a coding or data analysis competition that you may be interested in. Who knows, you might win some cash.



Getting Started ✕

Results

Recently Launched ▾

Store Sales - Time Series Forecasting
Use machine learning to predict grocery sales
Getting Started · Ongoing
Knowledge

I'm Something of a Painter Myself
Use GANs to create art - will you be the next Monet?
Getting Started · Code Competition · Ongoing
Prizes

Contradictory, My Dear Watson
Detecting contradiction and entailment in multilingual text using TPUs
Getting Started · Code Competition · Ongoing
Prizes

Petals to the Metal - Flower Classification on TPU
Getting Started with TPUs on Kaggle!
Getting Started · Code Competition · Ongoing
Knowledge

The GAN project is pretty neat!

More Kaggle...

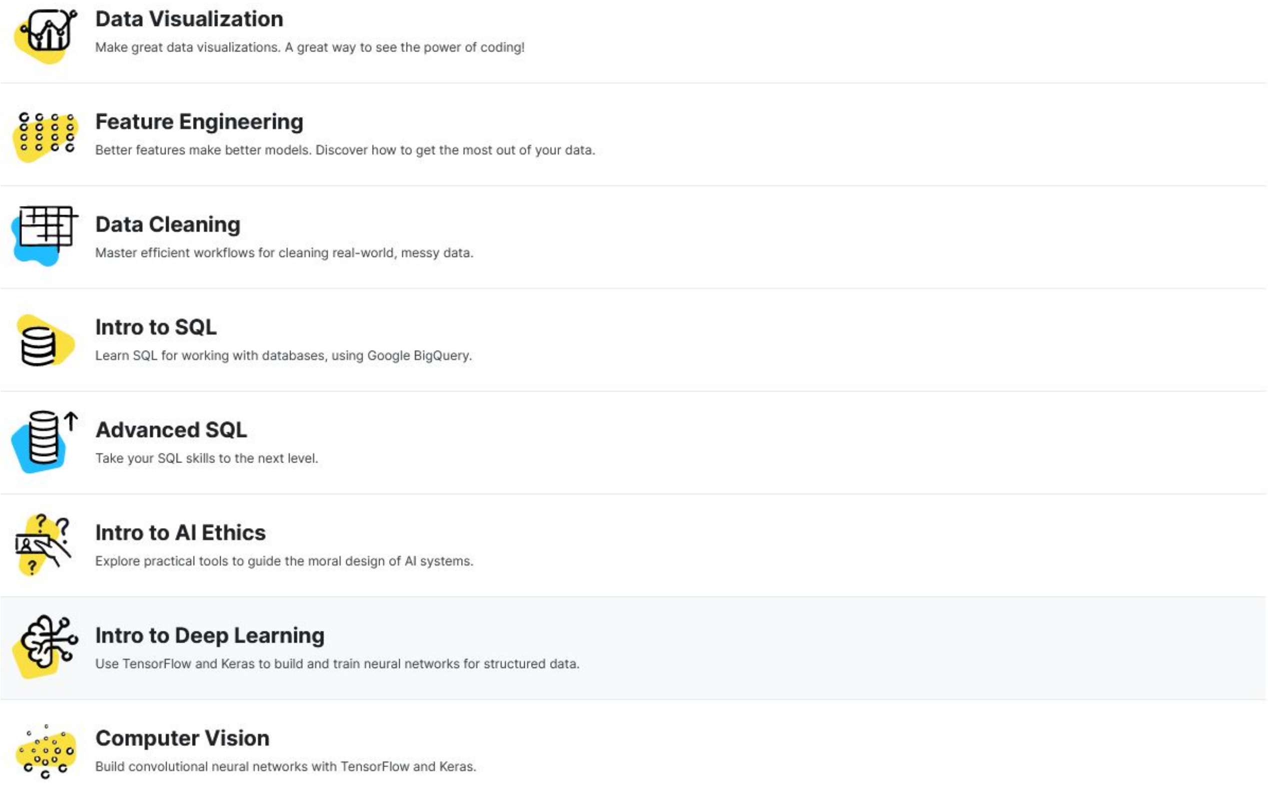
Courses

Gain the skills you need to do independent data science projects.

We pare down complex topics to their key practical components, so you gain usable skills in a few hours (instead of weeks or months).

The courses are **free**, and you can now earn certificates.

Course Offerings



Data Visualization

Make great data visualizations. A great way to see the power of coding!

Feature Engineering

Better features make better models. Discover how to get the most out of your data.

Data Cleaning

Master efficient workflows for cleaning real-world, messy data.

Intro to SQL

Learn SQL for working with databases, using Google BigQuery.

Advanced SQL

Take your SQL skills to the next level.

Intro to AI Ethics

Explore practical tools to guide the moral design of AI systems.

Intro to Deep Learning

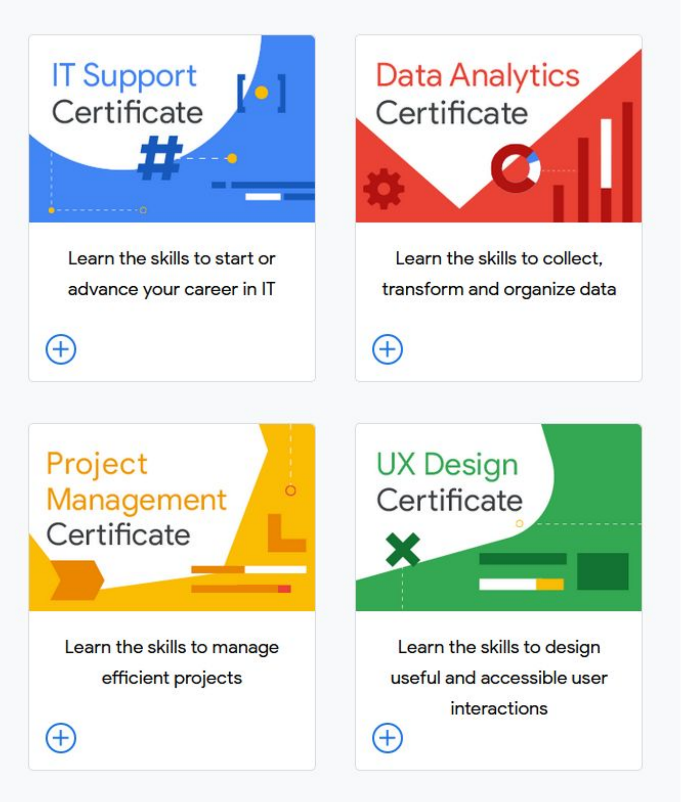
Use TensorFlow and Keras to build and train neural networks for structured data.

Computer Vision

Build convolutional neural networks with TensorFlow and Keras.

Grow With Google

No clue how good these are, but Data Analytics is an offering and if you sign up with your student email, you get one course a year for free on coursera. I would assume the content is fairly good if Google put it out.



IT Support Certificate

Learn the skills to start or advance your career in IT

Data Analytics Certificate

Learn the skills to collect, transform and organize data

Project Management Certificate

Learn the skills to manage efficient projects

UX Design Certificate

Learn the skills to design useful and accessible user interactions

Going farther

https://www.codecademy.com/catalog/language/python - Good resource for projects and learning

https://skulpt.org/ - Python in Browser

https://www.w3schools.com/python/default.asp - Information/documentation

https://www.learnpython.org/en/Hello%2C_World%21 - Interactive tutorials

https://towardsdatascience.com/ - Data Science Specifics

https://grow.google/dataanalytics/#?modal_active=none - Google data analytics certification

Hands On (Beginner)

If you have never used Python:

https://www.learnpython.org/en/Hello%2C_World%21

Demo

Analyzing COVID data with Pandas and Python

Spyder: https://www.spyder-ide.org/

Data:
'https://raw.githubusercontent.com/nytimes/covid-19-data/master/us-counties.csv'

'https://raw.githubusercontent.com/nytimes/covid-19-data/master/us-states.csv'

Add us on Social Media!

Twitter: **@ualbanyCDO**

Instagram: **ualbany_cdo**

Website: **uacyber.org**

Myinvolvement: **Cyber Defense Org**

We have a discord!