Feb 1st 2025

Data Science as a field. "Jane Wall"

- Data Science applications skills
- → electives to be needed.
- Data Science - reproducible analysis.
- Visualizing.
- Analysing.
- pitfalls
  - bias., sources, mitigates.
  - ethics.
- communicating results.

Data Science - Origin.

- our ability to count and categorize things
- capacity of data storage has grown
- capacity to process data has grown.
- DS as a field has grown.

Fields - Dept, Music, clicks, Purchases
Surveys - Analyse all of it.

Data science Venn diagram

Variety of jobs in Data Science

- getting and cleaning data
- Visualization and Analysis
- Shiny apps / Dashboards
- Modeling & Prediction.
- Machine Learning, AI
- Natural language processing
- Ethics & privacy, Policy.

How does Data Scientists spend their time

- Data Loading. — 19%
- Data cleansing — 26%
- Data Visualization — 21%
- Model selection — 11%
- Model training and scoring — 12%
- Deploying Models. — 11%

—

- Truth from data.
- No perfect answer.

Applications for machine learning.

Data Science.
feature engineering.
KS.
Billions of images. — scale.
data analysis - scale.

What does Google look for when they hire DS
- strong in one area.
  or
- balanced.
- able to explain the concepts in layman terms.

* product development lifecycle.

Data Management
Data wrangling.

Survey — 2000 responses.

R package

Recommender System -
Spacial Data Scientist
Data Science on a Map.

RStudio → IDE GUI

*.Rmd (R markdown)

>getwd()
-Rscript → *.R

Introduction to R - Basics

Basic Math

- much like calculator.
- Remainder of division.
  5 %% 2 = 1
- # This is comments.
- Assignment operator "<-"
  variable <- 100.
- Variable starts with small letters
- Functions starts with Capital letters

R Data type

- Numeric
  - floating point.
  - Integer.
- Logical
  - Boolean. TRUE/FALSE or T/F.
- Character.
- Class - Same as type in python.
  - class(TRUE) = logical.

- Vector - 1d array.
  - created using combine function. (c)
  - hvec <- c(1,2,3,4,5)
  - class(hvec) => numeric.

> cvec ← c('U','S','A')
> class(cvec) ⇒ character.
- Cannot mix data types in vectors.
  > v ← c(TRUE, 20, 40)
  > ~~class~~ v ⇒ (1, 20, 40)
  > class(v) ⇒ numeric.
- another example.
  > cvec ← c('USA', 20, 40)
  > CVEC ⇒ 'USA', '20', '40'.
  > class(cvec) ⇒ character.
- temps ← (72, 71, 68, 73, 69, 75, 76)
  > temps ⇒ 72, 71, 68 73 69 75 76
> names(temps) ← c('Mon', 'Tue', 'Wed', ...)
→ temps ⇒

Mon	Tue	Wed	...
72	71	68	

> days ← ('Mon' 'Tue' 'Wed' ...)
> names(temps) ← days.
> temps

Mon	Tue	Wed	...
72	71	68	

Vector operations.
> created vectorOperations.R under MAC

Vector Indexing and Slicing.
→ indexing starts from 1

Stackoverflow

Getting help.
> help("vector")
> ??vector.
> help.search("vector")

Dataframe
> data() ⇒ all available dataframe in R
> head(state.x77) ⇒ display first few rows
> tail(state.x77)
> str(state.x77) ⇒ structure of dataframe
> summary(state.x77)

Data Manipulation
> Dplyr package - Manipulating
> Tidyr package - cleaning data!
> Pipe operator %>%

Data visualisation - ggplot2.
- Built on the idea of adding layers
- First 3 layers are
  - Data - ggplot(dataframe)
  - Aesthetics - columns/features
  - Geometries. - geom_point()
  - Facets - multiple plots
  - Statistics - stat_smooth()
  - Coordinates
  - theme

Visualisation
- Coordinates and - resize ~~of~~ on Coordinate
- faceting - have multiple plots.

Bias
- feeling.
- How are you getting survey.
- Outliers. - ignore them? Interesting data.
- ML algorithm.
- Recognize bias - mitigate

Ethics → Why?
- It impacts many people or everyone.
- facial recognition -
- search on web browser
- ethical issue related to data
- ethical issues related ~~model~~ results from analysis and modeling that data using ML.
- ethical issues within Data Science.

NYC shooting data

incident_key dbl.
occur_date <chr>
occur_time <S3:hms>
BORO - <chr>

28,562 rows. 21 Columns

When records are read using read_csv, above details were obtained.

occur_date ✓
occur_time ✓
murder_flag ✓
perpetrator's age
sex
race.
victim's age
sex
race

→ 3 profiles.
perpetrator's profile
victim's profile
location profile.

Reproducible Research

Why? ① Makes your conclusion carry more weight
② Others can build off your work
③ Do it for your future self

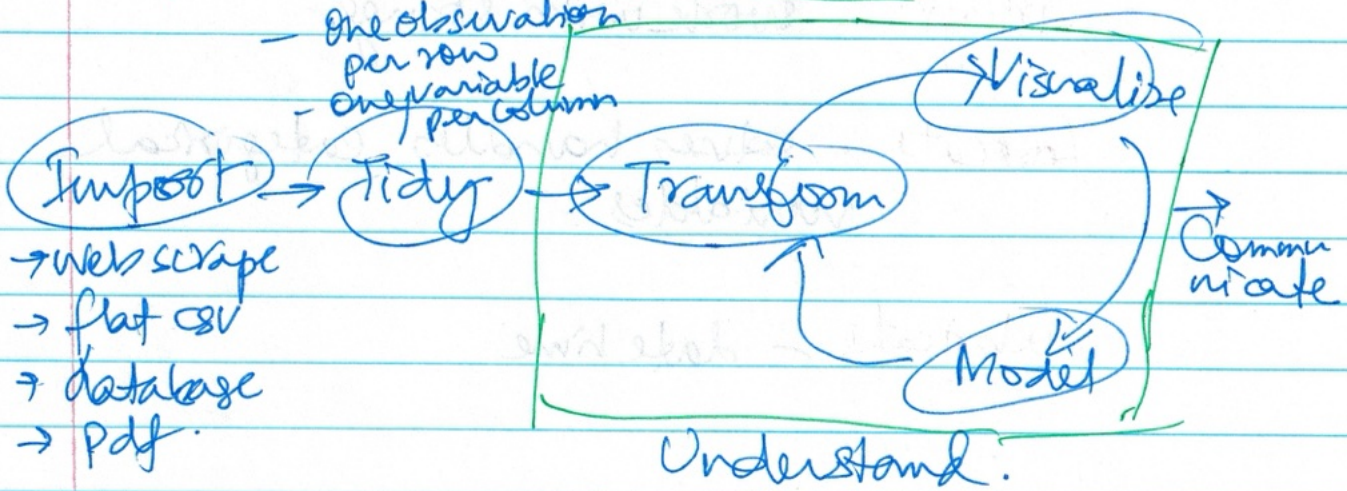
How? ① Document everything
- start with question and raw data

② All files are text files
- platform independence
- future proof

③ Make code readable
- comments
- formatting (indentation)
- meaningful names.
- each script does one task.

④ Use relative path.

Steps in data science process



Tidyverse packages

```
> install.packages("tidyverse")
> library(tidyverse)
```

ggplot2 - graphics

dplyr - data manipulation

tidyr - tidy data

readr - read rectangular data

purrr - enhances R's functional programming

tibble - modern re-imagining of the data frame

stringr - work with strings

forcats - solves handles categorical variables

lubricate - date time

Covid

> library(tidyverse)
> urls ← str_c(url_in, filenames)
↳ vector of urls.
> create dataframe variables
global_cases ← read_csv(url[1])

> different ways to explore datasets.
just print
head()
str()
summary() - summary by column/variable. (very useful)

> use of pivot_longer - to convert columns into single column, and values to another column.
name_to =
values_to.

pivot_longer - reshape data from a wide format to a long format.

syntax:

pivot_longer(data, cols, names_to, values_to)
data - data frame
cols - columns you want to pivot
This specifies which columns should be collapsed into key-value pairs.

-names_to - name of new column
-values_to- values

> then join data. (death data and cases data)
> global <- global_cases %>%
full_join(global_deaths) %>%
rename(Country_Region = "Country/Region",
Province_State = "Province/State") %>%
mutate(date = mdy(date))

mutate - is used to create new column or modify existing coln.

example - new column.
df contains x and y
df <- df %>%
mutate(z = x+y) creates z

example - modify existing column.
df <- df %>%
mutate(y = y^2).

library(lubridate) - for using
date = mdy(date)

> global <- global %>% filter(cases > 0)

> then analyse data and then visualize it.
  > group by is used for categorical data.
  > Summarize - used for creating summary statistics.

> visualisation
  Covid 19 in "US"
  Covid 19 in "new York"
  new cases..

> use of lag() function ⇒ - used to create a lagged version of a vector, shifting elements by a specified number of positions. It is used in time series analysis to compare current values with past values.

> exercise.
  - take the case of different country.
  - worst and best state.

> linear model.
  - lm.

Bias

Add bias in the project report.

NYPD shooting

- Complete all the steps in data science process in reproducible manner.
- look at NYPD shooting data and analyse
- grading
  - Data science process
    - Reproducibility.
    - clear writing
    - validity of analysis
    - identification of bias.
- Rubric

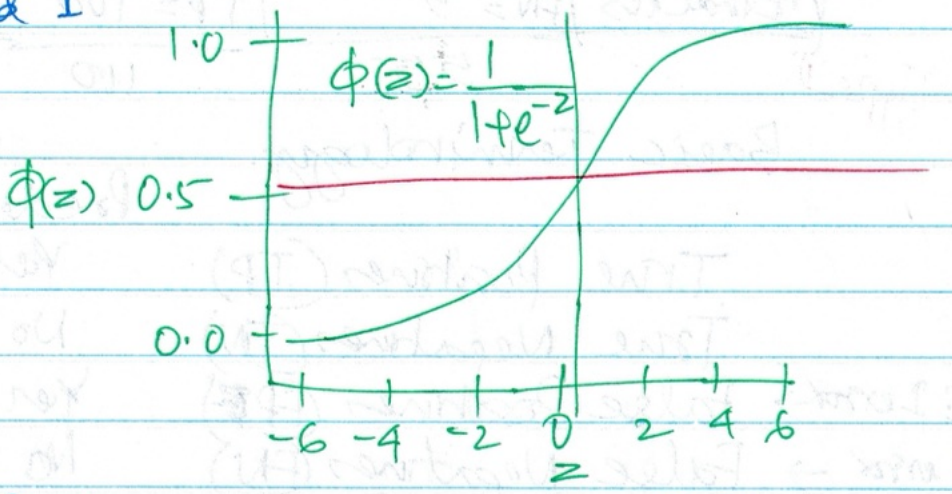
✓ import	Reproducible.
✓ tidy	visualisation
✓ analyse.	analysis.
✓ 2 visualisation	

  - ~~bias~~ bias
  - ~~1~~ model.
  - upload .html / pdf.

> plots are good way to find correlation

Logistic Regression

> the sigmoid (aka Logistic) Function takes in any value and outputs it to be between 0 and 1



$\phi(z)$ is always between 0 and 1.

Linear Model $y = b_0 + b_1 x$ ⟹ transforms into

Logistic Model $p = \frac{1}{1+e^{-(b_0+b_1 x)}}$

> we usually set a cutoff point (say 0.5) anything below probability is zero anything above probability is (1) One.

Evaluation of classification models

> Confusion Matrix.

n = 165	PREDICTED NO	PREDICTED YES	
Actual NO	TN = 50	FP = 10	60
Actual YES	FN = 5	TP = 100	105
	55	110	

Type I

Type II

Basic Terminology.

	Predicted	Actual
True Positives (TP)	Yes	Yes.
True Negatives (TN)	No	No
TYPE I error → False Positives (FP)	Yes	No
TYPE II error → False Negatives (FN)	No	Yes.

Accuracy — how often it is right

$$\text{Correct} = \frac{TP + TN}{\text{Total}} = \frac{100+50}{165} = 91\%$$

Misclassification Rate — How often it is wrong.
(error rates).

$$\text{Misclassification} = \frac{FP + FN}{\text{Total}} = \frac{15}{165} = 9\%$$

Communicating Results

→ Reports
→ Presentations
→ Elevator Pitch.

Written report -

- be organized
- Easy to read - tell the story
- key visualizations
- Format, others can open
- good grammer / sentence structure
- Why should I care?
- Where is the data from.
- What does it tell you - analysis and visualization
- What do you conclude.
- How could you be wrong.

Presentations.

- don't be boring
- be organized
- tell the story
- key visualizations
- Make eye contact (or camera)

Presentations (DONT)

- Dont read slides.

- Don't have lots of words on slide
- Don't be monotone
- Don't go over time limit.
- Don't overwhelm with many similar visualizations.

Presentations (DO)

- Include pictures. to make your main point
- target talk to audience
- be enthusiastic
- practice
- maintain eye contact
- Use whiteboard or other options

Elevator pitch

- explain to non-technical person.
- target is 13 year old.
- short approx: 1 minute.

Build portfolio

→ github

Meetup

→ Data Science

25th Feb

Sequence

Question of Interest
- Import Libraries
- Import data.
- Read global da & US data.

Explore Columns

Tidy and Transform Global data.

Tidy and Transform US data.

Visualization of US data.
- US
- state - New York
- State - Arizona
- New cases/deat US
- New ~~cases~~ deaths_ state
- Best state
- worst state

Bias Sources

Political affiliation
Population density.
extent of lockdown.
Climate of area.