sharing meaningful shares

by dasapta erwin irawan ORCID (http://orcid.org/0000-0002-1526-0863)

presented at iccset 2018 >> slides here (https://github.com/dasaptaerwin/iccset2018)

a quick intro

- everything now is labelled with **i** or **smart**
- but people are forgetting something called **share**
- sure ... everyone shares but lack of **meaningful share**
- the available rooms for feedback are only for **likes** and **shares**

so what is a meaningful share (scientifically)

something that can be freely:

- downloaded and read **(this where most of us stop)**
- verified and reproduced
- remixed and reused

now let's have an example

- suppose i knew nothing about doing linear regression using python
- so i search online and found this page >> https://github.com/justmarkham/DAT4/blob/master/notebooks/08_linear_regression.ipynb (https://github.com/justmarkham/DAT4/blob/master/notebooks/08_linear_regression.ipynb)
- so what would i do next?

- i would visit and read the page,
- i would verify and simply reproduce the code by copy and pasting the cells
  - or just download it and run the notebook
- the if it's running ok, then i would apply it to my own dataset

here's a walkthrough demo

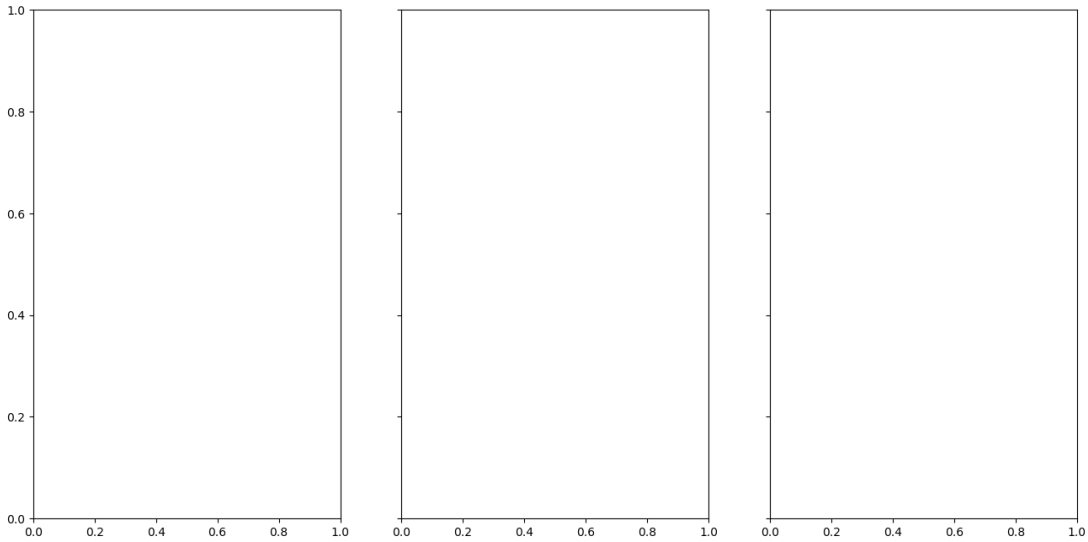
In [43]:
```
# https://github.com/justmarkham/DAT4/blob/master/notebooks/08_linear_regression.ipynb
# imports
import pandas as pd
import matplotlib.pyplot as plt

# this allows plots to appear directly in the notebook
%matplotlib inline

# read data into a DataFrame
data = pd.read_csv('http://www-bcf.usc.edu/~gareth/ISL/Advertising.csv', index_col=0)
data.head()
```

Out[43]:

	TV	radio	newspaper	sales
1	230.1	37.8	69.2	22.1
2	44.5	39.3	45.1	10.4
3	17.2	45.9	69.3	9.3
4	151.5	41.3	58.5	18.5
5	180.8	10.8	58.4	12.9

In [40]:
```
# visualize the relationship between the features and the response using scatte
rplots
fig, axs = plt.subplots(1, 3, sharey=True)
data.plot(kind='scatter', x='TV', y='Sales', ax=axs[0], figsize=(16, 8))
data.plot(kind='scatter', x='Radio', y='Sales', ax=axs[1])
data.plot(kind='scatter', x='Newspaper', y='Sales', ax=axs[2])
```

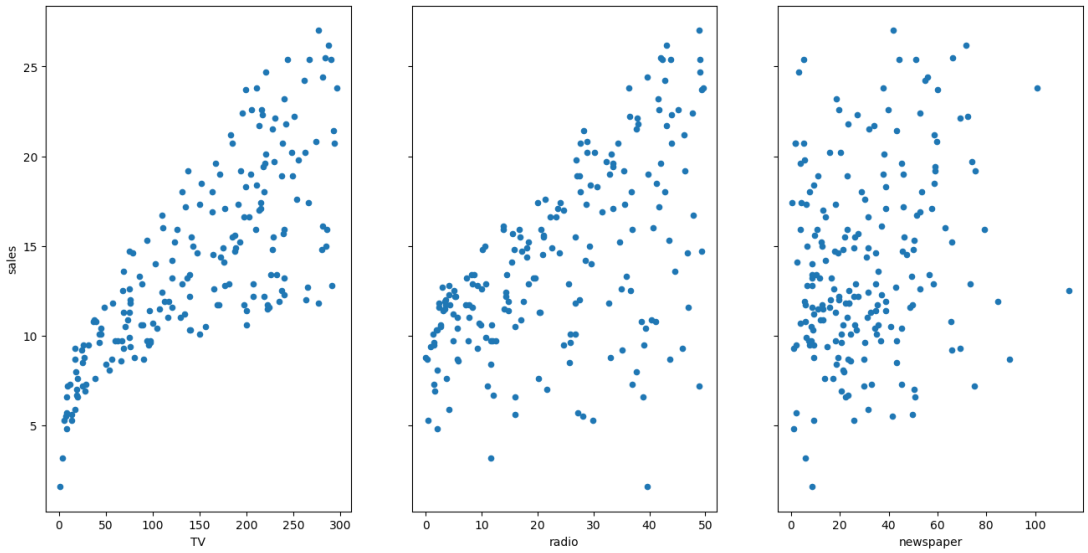
```
---------------------------------------------------------------------------
KeyError                                  Traceback (most recent call last)
/Applications/anaconda/lib/python3.4/site-packages/pandas/indexes/base.py in g
et_loc(self, key, method, tolerance)
   2133             try:
-> 2134                 return self._engine.get_loc(key)
   2135             except KeyError:

pandas/index.pyx in pandas.index.IndexEngine.get_loc (pandas/index.c:4433)()

pandas/index.pyx in pandas.index.IndexEngine.get_loc (pandas/index.c:4279)()

pandas/src/hashtable_class_helper.pxi in pandas.hashtable.PyObjectHashTable.ge
t_item (pandas/hashtable.c:13742)()

pandas/src/hashtable_class_helper.pxi in pandas.hashtable.PyObjectHashTable.ge
t_item (pandas/hashtable.c:13696)()

KeyError: 'Sales'

During handling of the above exception, another exception occurred:

KeyError                                  Traceback (most recent call last)
<ipython-input-40-0d6954da619f> in <module>()
      1 # visualize the relationship between the features and the response usi
ng scatterplots
      2 fig, axs = plt.subplots(1, 3, sharey=True)
----> 3 data.plot(kind='scatter', x='TV', y='Sales', ax=axs[0], figsize=(16,
8))
      4 data.plot(kind='scatter', x='Radio', y='Sales', ax=axs[1])
      5 data.plot(kind='scatter', x='Newspaper', y='Sales', ax=axs[2])

/Applications/anaconda/lib/python3.4/site-packages/pandas/tools/plotting.py in
__call__(self, x, y, kind, ax, subplots, sharex, sharey, layout, figsize, use_
index, title, grid, legend, style, logx, logy, loglog, xticks, yticks, xlim, y
lim, rot, fontsize, colormap, table, yerr, xerr, secondary_y, sort_columns, **
kwds)
   3772                           fontsize=fontsize, colormap=colormap, table=
table,
   3773                           yerr=yerr, xerr=xerr, secondary_y=secondary_
y,
-> 3774                           sort_columns=sort_columns, **kwds)
   3775     __call__.__doc__ = plot_frame.__doc__
   3776

/Applications/anaconda/lib/python3.4/site-packages/pandas/tools/plotting.py in
plot_frame(data, x, y, kind, ax, subplots, sharex, sharey, layout, figsize, us
e_index, title, grid, legend, style, logx, logy, loglog, xticks, yticks, xlim,
ylim, rot, fontsize, colormap, table, yerr, xerr, secondary_y, sort_columns,
**kwds)
   2641                  yerr=yerr, xerr=xerr,
   2642                  secondary_y=secondary_y, sort_columns=sort_columns,
-> 2643                  **kwds)
   2644
   2645

/Applications/anaconda/lib/python3.4/site-packages/pandas/tools/plotting.py in
_plot(data, x, y, subplots, ax, kind, **kwds)
   2468         plot_obj = klass(data, subplots=subplots, ax=ax, kind=kind, **
kwds)
   2469
-> 2470     plot_obj.generate()
   2471     plot_obj.draw()
   2472     return plot_obj.result

/Applications/anaconda/lib/python3.4/site-packages/pandas/tools/plotting.py in
generate(self)
   1041         self._compute_plot_data()
   1042         self._setup_subplots()
-> 1043         self._make_plot()
```



oops error message and what should we do? we could check and make it right

In [41]:
```
# visualize the relationship between the features and the response using scatterplots
fig, axs = plt.subplots(1, 3, sharey=True)
data.plot(kind='scatter', x='TV', y='sales', ax=axs[0], figsize=(16, 8))
data.plot(kind='scatter', x='radio', y='sales', ax=axs[1])
data.plot(kind='scatter', x='newspaper', y='sales', ax=axs[2])
```

Out[41]: <matplotlib.axes._subplots.AxesSubplot at 0x1151da390>



- now let us replicate: apply those codes to our dataset
  - i used my own dataset here: https://github.com/dasaptaerwin/nutrient2018 (https://github.com/dasaptaerwin/nutrient2018)
  - i copied the dataset into a new working folder

In [20]:
```
# read data into a DataFrame
data = pd.read_csv('data/data_dago_atas.csv', index_col=0)
data.head()
```

Out[20]:

NO	COLLECTOR	DATE	DAY	WEEK	TIME_SLOT	LOCATION	TYPE	CI2	PH	KH	GH	N
1	Niki	2018-03-24	sat	END	1	dago_atas	Air Sungai	10.0	7.2	10	7	5.
2	Niki	2018-03-24	sat	END	1	dago_atas	Air Sumur	3.0	8.0	6	7	5.
3	Niki	2018-03-25	sun	END	1	dago_atas	Air Sungai	10.0	7.2	10	7	5.
4	Niki	2018-03-25	sun	END	1	dago_atas	Air Sumur	3.0	8.0	6	7	5.
5	Niki	2018-03-31	sat	END	1	dago_atas	Air Sungai	10.0	8.0	10	7	5.

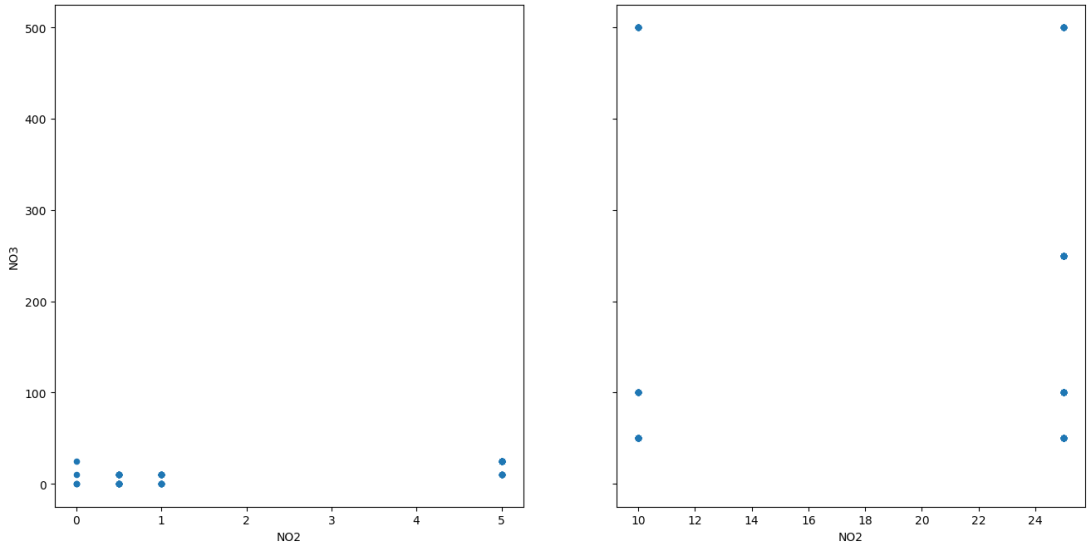
In [21]:
```
# read data into a DataFrame
data2 = pd.read_csv('data/data_braga.csv', index_col=0)
data.head()
```

Out[21]:

NO	COLLECTOR	DATE	DAY	WEEK	TIME_SLOT	LOCATION	TYPE	CI2	PH	KH	GH	N
1	Niki	2018-03-24	sat	END	1	dago_atas	Air Sungai	10.0	7.2	10	7	5.
2	Niki	2018-03-24	sat	END	1	dago_atas	Air Sumur	3.0	8.0	6	7	5.
3	Niki	2018-03-25	sun	END	1	dago_atas	Air Sungai	10.0	7.2	10	7	5.
4	Niki	2018-03-25	sun	END	1	dago_atas	Air Sumur	3.0	8.0	6	7	5.
5	Niki	2018-03-31	sat	END	1	dago_atas	Air Sungai	10.0	8.0	10	7	5.

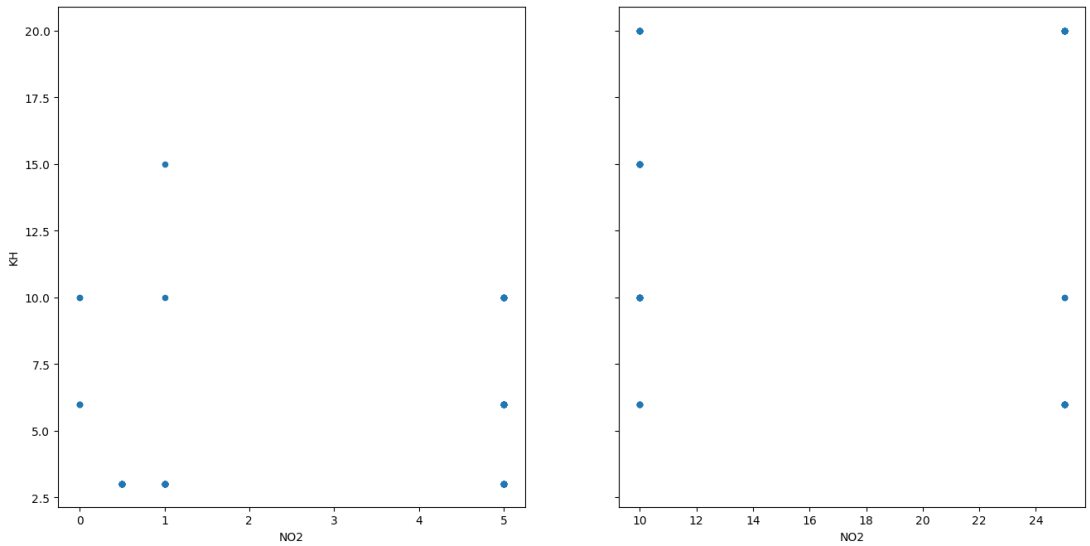
In [22]:
```
# visualize the relationship between the features and the response using scatterplots
fig, axs = plt.subplots(1, 2, sharey=True)
data.plot(kind='scatter', x='NO2', y='NO3', ax=axs[0], figsize=(16, 8))
data2.plot(kind='scatter', x='NO2', y='NO3', ax=axs[1])
```

Out[22]: <matplotlib.axes._subplots.AxesSubplot at 0x113e42b00>



In [23]:
```
# visualize the relationship between the features and the response using scatterplots
fig, axs = plt.subplots(1, 2, sharey=True)
data.plot(kind='scatter', x='NO2', y='KH', ax=axs[0], figsize=(16, 8))
data2.plot(kind='scatter', x='NO2', y='KH', ax=axs[1])
```

Out[23]: <matplotlib.axes._subplots.AxesSubplot at 0x114286828>



... and how we do that?

make it easy to access

- **open-non profit** venues (repositories) not closed-`not for profit` venues (please not RG, Academia),

- **no sign ups** necessary
- no hidden **monetization** scheme

make it easy to use

- in **text** format (ascii, odt, docx), prefereably not pdf
- **raw** data in text not table in pdf format
- clear **documentation**

make it legal

- assign **proper** license
- preferably **CC-based** licenses or the equivalent
- attached **data agreement** or data usage guidelines

these are some examples

project repository

- we use gitlab (gitlab.com/derwinirawan), github (github.com/dasaptaerwin), and OSF (https://osf.io/he3j7/) to host our live projects >> view some examples (links will be added) Nutrient2018 (https://gitlab.com/derwinirawan/nutrient_2018), pub analytics (https://github.com/dasaptaerwin/pubanalytics), literate programming (https://github.com/dasaptaerwin/literateprogrammingSNIPS2018), preliminary mapping cikapundung (https://osf.io/g5fex/).
- we share as we're working on them >> view some examples (links will be added) sharing proposal (https://derwinirawan.wordpress.com/2018/10/09/hidrogeologi-kawasan-pemakaman-umum/), live slide decks (http://dasaptaerwin.net/wp/2018/08/terbuka-atau-tertinggal-paparan-rakernas-rji-agustus-2018.html).
- we systematically set our files following this folders:
  - data
  - code
  - output
  - report

- we use free-opensource apps or at least choose your apps wisely that most people can re-do your work using theirs:
  - r (https://cran.r-project.org)
  - python (https://anaconda.org/anaconda/python)
  - LaTeX (overleaf.com), Markdown (https://daringfireball.net)
  - etc

- invite visitor to leave feedback:
  - use free-open service like Hypothes.is (https://web.hypothes.is/)

- on the other hand, make time to drop some comments to others

why we're doing this

sharing is caring

increase the benefit of your science to society

you're all i-t people, you should be long aware that this internet thingy will help you increase your impact ... a lot

so why not doing it?

take home message

... please, making **impact** is way beyond chasing **H-index** and **Journal Impact Factor**

a bit about me

this is where i work:

applied geology research group (https://medium.com/open-science-indonesia/trend-pengelolaan-jurnal-di-era-digital-ee1564423251),

faculty of earth sciences and technology (http://www.fitb.itb.ac.id/), institut teknologi bandung

this is my passion: INArxiv (inarxiv.id)

i blog about my:

- experience in learning (dasaptaerwin.net)
- real work in hydrogeology (derwinirawan.wordpress.com)
- passion in open science indonesia (https://medium.com/open-science-indonesia)

and like any **normal** human beings ... i tweet ... from @dasaptaerwin (twitter.com/dasaptaerwin)

so share a meaningful share

`#sainsterbuka` `#openscience`

`#terbukaatautertinggal` `#beopenorleftbehind`