NCRM National Centre for Research Methods

An introduction to Jupyter Notebooks for Social Science Research

Vernon Gale

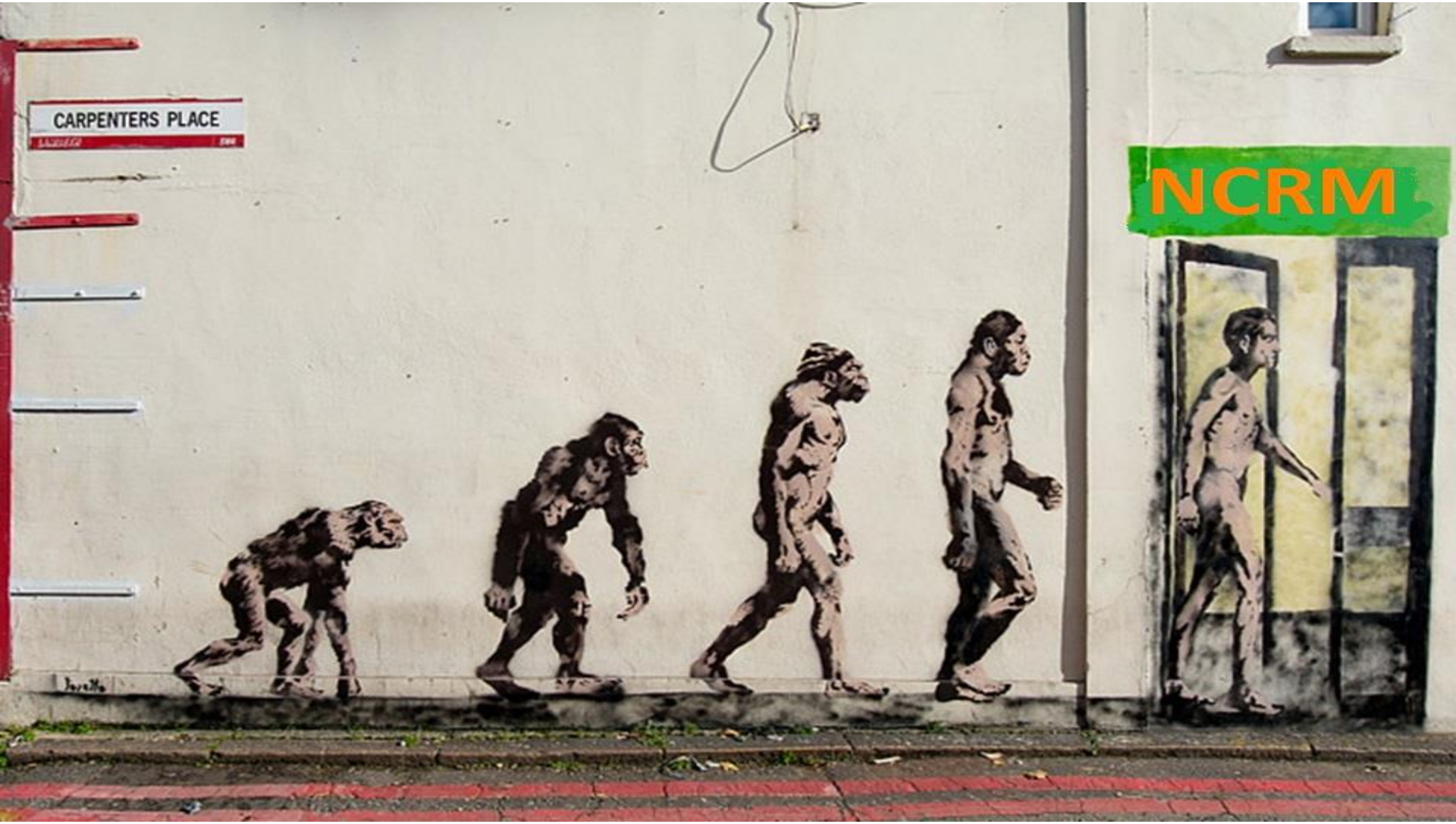
University of Edinburgh

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL

UNIVERSITY OF Southampton

MANCHESTER 1824 The University of Manchester

THE UNIVERSITY OF EDINBURGH


CARPENTERS PLACE
NCRM


The Workflow of Data Analysis Using Stata
J. SCOTT LONG

scarc.library.oregonstate.edu/coll/pauling/rnb/index.html

As with many scientists, Linus Pauling utilized bound notebooks to keep track of the details of his research as it unfolded. A testament to the remarkable length and diversity of Dr. Pauling's career, the Pauling Papers holdings include forty-six research notebooks spanning the years of 1922 to 1994 and covering any number of the scientific fields in which Dr. Pauling involved himself. In this regard, the notebooks contain many of Pauling's laboratory calculations and experimental data, as well as scientific conclusions, ideas for further research and numerous autobiographical musings.

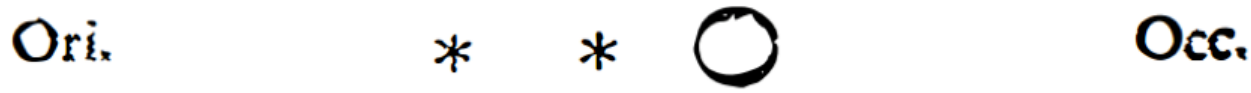
Research Notebook 01
1922
Research Notebook 02
1922-1923, 1932, 1934, 1936, 1973, 1985
Research Notebook 03
1923-1925
Research Notebook 04
1923-1924, 1928-1930
Research Notebook 05

Research Notebook 13
1935-1936, 1938-1939
Research Notebook 14
1936-1939, 1949, 1952
Research Notebook 15
1935, 1937, 1968
Research Notebook 16
1935-1956
Research Notebook 17
1939-1941, 1971, 1988

Research Notebook 24
1953, 1956, 1962, 1963, 1967, 1968, 1969, 1970, 1973
Research Notebook 25
1958, 1964-1966
Research Notebook 26
1955, 1964-1969, 1974-1976, 1980-1982, 1987, 1990-1991
Research Notebook 27
1952-1954, 1960-1961, 1964, 1971-

Research Notebook 35b
1938-1939, 1946, 1955, 1968, 1986-1988
Research Notebook 36
1980-1981, 1986-1987
Research Notebook 37
1971, 1983
Research Notebook 38
1980-1981, 1983, 1985, 1989
Research Notebook 39

But on January 10th the stars appeared in the following position with regard to Jupiter; there were two only, and both on the east side



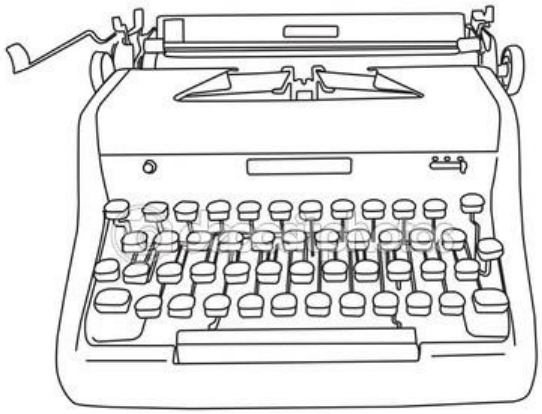
of Jupiter, the third, as I thought, being hidden by the planet.

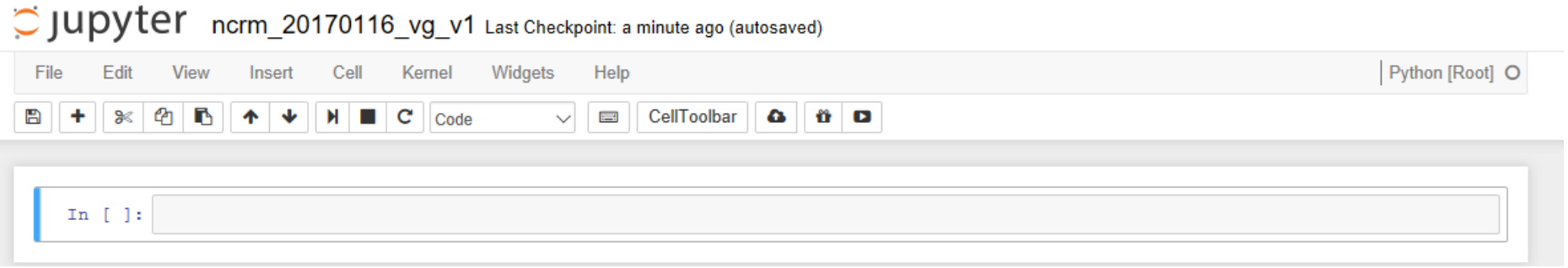

jupyter

https://www.youtube.com/watch?v=BmHPoBpZoJ4


jupyter

Documentation; Portability; Language agnostic; Rich visual outputs; Big data tools; Teaching tools; Collaboration




jupyter
ncrm_20170116_vg_v1
Last Checkpoint: a minute ago (autosaved)
File
Edit
View
Insert
Cell
Kernel
Widgets
Help
Python [Root]
Code
CellToolbar
In [ ]:

In [4]:
```
summarize
```

In [4]:
```
summarize
```

```
    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
        case |      1,580    517.7411    284.8605         1       1003
        femp |      1,580    .6455696    .4784918         0          1
        mune |      1,580    .0740506    .2619362         0          1
        time |      1,580         7.2    3.981019         0         13
        und1 |      1,580    .0746835    .2629633         0          1
-------------+---------------------------------------------------------
        und5 |      1,580    .2974684    .4572891         0          1
         age |      1,580    36.01013    9.114841        18         60
```

In [4]:
```
summarize
```

```
-------------+--------------------------------------------------------
        case |      1,580    517.7411    284.8605          1       1003
       femp |      1,580    .6455696    .4784918          0          1
       mune |      1,580    .0740506    .2619362          0          1
       time |      1,580         7.2    3.981019          0         13
       und1 |      1,580    .0746835    .2629633          0          1
-------------+--------------------------------------------------------
       und5 |      1,580    .2974684    .4572891          0          1
        age |      1,580    36.01013    9.114841         18         60
```

The data mirror a real example of data analysed in Davies et al. (1992).

The dataset is a panel of 155 married women.

Davies, Richard B., Peter Elias, and Roger Penn. "The relationship between a husband's unemployment and his wife's participation in the labour force." Oxford Bulletin of Economics and Statistics 54.2 (1992): 145-171.

In [11]:
```
logit femp mune und5
```

```
Iteration 0:   log likelihood = -1027.2309
Iteration 1:   log likelihood = -879.88806
Iteration 2:   log likelihood = -878.68101
Iteration 3:   log likelihood = -878.67998
Iteration 4:   log likelihood = -878.67998

Logistic regression                             Number of obs     =      1,580
                                                LR chi2(2)        =     297.10
                                                Prob > chi2       =     0.0000
Log likelihood = -878.67998                     Pseudo R2         =     0.1446

------------------------------------------------------------------------------
        femp |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        mune |  -1.703308   .2358489    -7.22   0.000    -2.165563   -1.241053
        und5 |  -1.733521   .1221909   -14.19   0.000    -1.973011   -1.494031
       _cons |   1.306829   .0744154    17.56   0.000     1.160978    1.452681
------------------------------------------------------------------------------
```

In [3]:
```
mylogit <- glm(femp ~ mune + und5, data = mydata, family = "binomial")

summary(mylogit)
```

```
Call:
glm(formula = femp ~ mune + und5, family = "binomial", data = mydata)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.7586  -1.0024   0.6922   0.6922   2.1177

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.30683    0.07442  17.561  < 2e-16 ***
mune        -1.70331    0.23585  -7.222 5.12e-13 ***
und5        -1.73352    0.12219 -14.187  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2054.5  on 1579  degrees of freedom
Residual deviance: 1757.4  on 1577  degrees of freedom
AIC: 1763.4
```

In [6]:
```
independentVar = ['mune', 'und5', 'Int']
logReg = sm.Logit(df['femp'] , df[independentVar])
answer = logReg.fit()
```

```
Optimization terminated successfully.
         Current function value: 0.556127
         Iterations 5
```

the results are in the oject "answer"

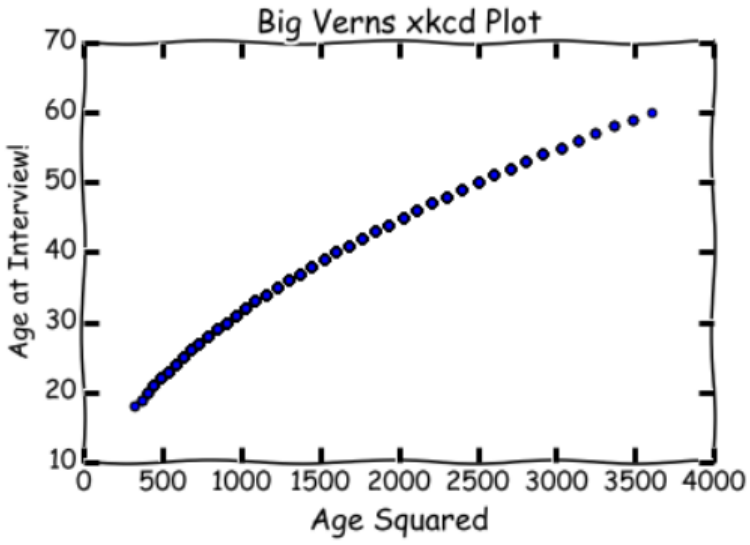
In [9]:
```
answer.summary()
```

Out[9]:

Logit Regression Results

Dep. Variable:	femp	**No. Observations:**	1580
Model:	Logit	**Df Residuals:**	1577
Method:	MLE	**Df Model:**	2
Date:	Fri, 14 Oct 2016	**Pseudo R-squ.:**	0.1446
Time:	10:13:23	**Log-Likelihood:**	-878.68
converged:	True	**LL-Null:**	-1027.2
		LLR p-value:	3.056e-65

	coef	std err	z	P>|z|	[95.0% Conf. Int.]
mune	-1.7033	0.236	-7.222	0.000	-2.166 -1.241
und5	-1.7335	0.122	-14.187	0.000	-1.973 -1.494
Int	1.3068	0.074	17.561	0.000	1.161 1.453

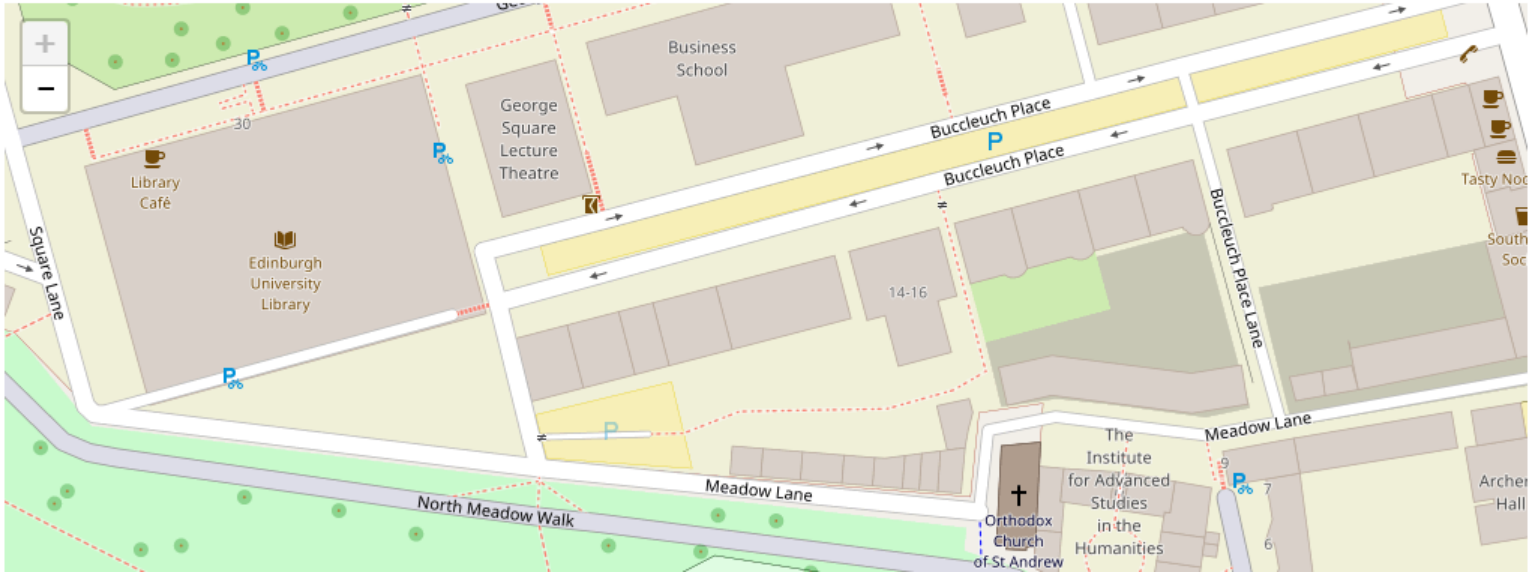


Another inventive use of the wemp dataset

Using an open street map

I've recently moved to a more commodious office in Buccleuch Place. Here is an example of an open source map on my new *hood*.

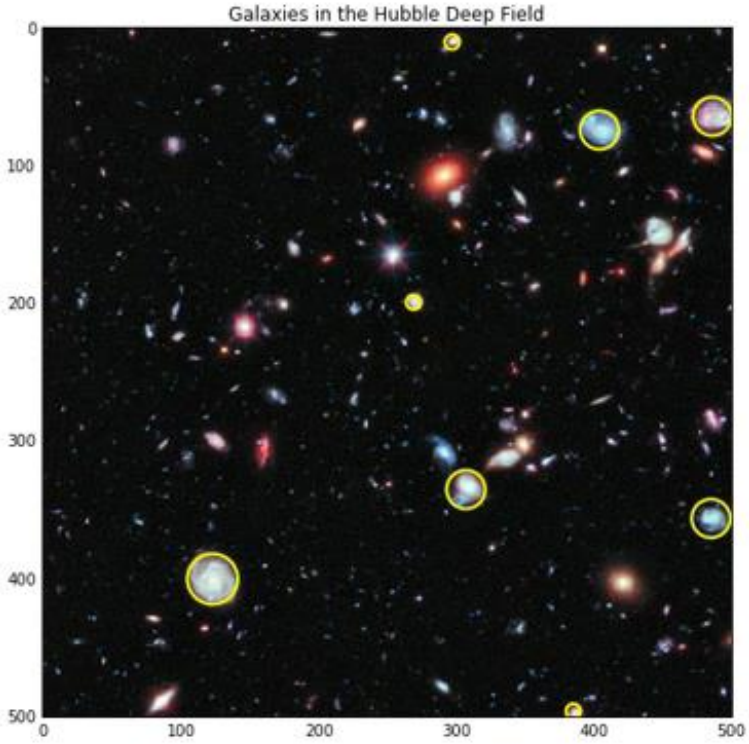
In [5]:
```
from ipyleaflet import Map
Map(center=[55.942535, -3.187269], zoom=20)
```



max_sigma 40

threshold 0.02

☐ gray



http://lorenabarba.com/

jupyter

Install About Resources Documentation NBViewer Widgets Blog Donate

Open source, interactive data science and scientific computing across over 40 programming languages.

https://jupyter.org/

NCRM National Centre for Research Methods

For more information visit

www.ncrm.ac.uk/resources/online