Προγραμματισμός με τη γλώσσα python

Alexandros Kanterakis kantale@ics.forth.gr (mailto:kantale@ics.forth.gr)

Διάλεξη 9η, Τρίτη 10 Δεκεμβρίου 2019

Το pandas (http://pandas.pydata.org/) είναι μία βιβλιοθήκη σε python για ανάλυση δεδομένων. Υιοθετεί τη φιλοσοφία της Matlab και R για οργάνωση 2-διάστατων δεδομένων σε μία ειδική δομή που ονομάζεται data frame. Στη βιοπληροφορική το pandas συνήθως είναι χρήσιμο για να κάνοθμε εργασίες που συνήθως γίνονται με το excel. Τα πλεονεκτήματα του pandas είναι:

- Πάρα πολύ γρήγορο. Είναι υλοποιημένο σε C (η python "τρέχει" από πάνω) και έχει πολύ καλή απόδοση για πίνακες που έχουν μέχρι και εκατομύτια από γραμμές.
- Παρέχει ένα interface το οποίο προσομοιάζει τις βάσεις δεδομένων. Με αυτόν τον τρόπο μπορούμε να γράφουμε σύντομες εκφράσεις που κάνουν πολύπλοκες διεργασίες.
- Υποστηρίζεται από τρίτες βιβλιοθήκες για visualization, Machine Learning (π.χ. sci-kit (http://scikit-learn.org/stable/)) και στατιστική (π.χ. statmodels (http://statsmodels.sourceforge.net/)).
- Παρέχει δικές του μεθόδους για γρήγορο plotting και στατιστική ανάλυση
- Εύκολη και γρήγορο input / output σε διάφορα formats (excel included)

Συνήθως κάνουμε import το pandas ως εξής:

In [1]:
```
import pandas as pd
```

Αν δεν υπάρχει εγκαταστημένων τότε μπορείτε να το εγκαταστείσετε ως εξής:

```
pip install pandas
```

Προσοχή. Πρέπει το `pip` να βρίσκεται στην ίδια τοποθεσία που βρίσκεται και η python

Για τη παρούσα διάλεξη θα χρησιμοποιήσουμε έναν κατάλογο από GWA studies (https://en.wikipedia.org/wiki/Genome-wide_association_study). Ο κατάλογος βρίσκεται σε αυτό το link: https://www.ebi.ac.uk/gwas/api/search/downloads/full (https://www.ebi.ac.uk/gwas/api/search/downloads/full) για να το κατεβάσετε τοπικά τρέξτε:

In [4]:
```
!wget -O gwas.tsv  "https://www.ebi.ac.uk/gwas/api/search/downloads/full"
```

```
--2016-12-15 14:50:41--  https://www.ebi.ac.uk/gwas/api/search/downloads/full
Resolving www.ebi.ac.uk... 193.62.193.80
Connecting to www.ebi.ac.uk|193.62.193.80|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: unspecified [text/tsv]
Saving to: ‘gwas.tsv’

gwas.tsv                [          <=>       ]  17.93M  3.90MB/s    in 5.3s

2016-12-15 14:50:47 (3.36 MB/s) - ‘gwas.tsv’ saved [18796575]
```

Η παραπάνω εντολή σώζει τον κατάλογο στο αρχείο: `gwas.tsv`

Για να το φορτώσουμε τρέχουμε:

In [3]:
```
gwas = pd.read_csv('gwas.tsv', sep='\t')
```

```
/Users/alexandroskanterakis/anaconda3/envs/arkalos/lib/python3.6/site-packages
/IPython/core/interactiveshell.py:2728: DtypeWarning: Columns (23,27) have mix
ed types. Specify dtype option on import or set low_memory=False.
  interactivity=interactivity, compiler=compiler, result=result)
```

Για να τυπώσουμε μία σύνοψη (πρώτες και τελευταίες γραμμές) των δεδομένων τρέχουμε:

In [4]:
```
gwas
```

Out[4]:

	DATE ADDED TO CATALOG	PUBMEDID	FIRST AUTHOR	DATE	JOURNAL	LINK	ST
0	2009-09-28	18403759	Ober C	2008-04-09	N Engl J Med	www.ncbi.nlm.nih.gov/pubmed/18403759	Effe variat CHI3l serum YK
1	2008-06-16	18369459	Liu Y	2008-04-04	PLoS Genet	www.ncbi.nlm.nih.gov/pubmed/18369459	A gen assoc stu psorias
2	2008-06-16	18385676	Amos CI	2008-04-03	Nat Genet	www.ncbi.nlm.nih.gov/pubmed/18385676	Genome- assoc scan (SNPs ic
3	2008-06-16	18385676	Amos CI	2008-04-03	Nat Genet	www.ncbi.nlm.nih.gov/pubmed/18385676	Genome- assoc scan (SNPs ic
4	2008-06-16	18385676	Amos CI	2008-04-03	Nat Genet	www.ncbi.nlm.nih.gov/pubmed/18385676	Genome- assoc scan (SNPs ic
5	2008-06-16	18385738	Hung RJ	2008-04-03	Nature	www.ncbi.nlm.nih.gov/pubmed/18385738	suscept locus fo cancer
6	2008-09-16	18385739	Thorgeirsson TE	2008-04-03	Nature	www.ncbi.nlm.nih.gov/pubmed/18385739	A v assoc with nic depender
7	2008-06-16	18372901	Tenesa A	2008-03-30	Nat Genet	www.ncbi.nlm.nih.gov/pubmed/18372901	Genome- assoc scan ider a c
8	2008-06-16	18372901	Tenesa A	2008-03-30	Nat Genet	www.ncbi.nlm.nih.gov/pubmed/18372901	Genome- assoc scan ider a c
9	2008-06-16	18372901	Tenesa A	2008-03-30	Nat Genet	www.ncbi.nlm.nih.gov/pubmed/18372901	Genome- assoc scan ider a c
10	2008-06-16	18372905	Tomlinson IP	2008-03-30	Nat Genet	www.ncbi.nlm.nih.gov/pubmed	A gen assoc

Για να τυπώσουμε μόνο κάποιες γραμμές:

In [5]:
```
gwas[0:3] # Πρώτες 3 γραμμές
```

Out[5]:

	DATE ADDED TO CATALOG	PUBMEDID	FIRST AUTHOR	DATE	JOURNAL	LINK	STUDY	DISE
0	2009-09-28	18403759	Ober C	2008-04-09	N Engl J Med	www.ncbi.nlm.nih.gov/pubmed/18403759	Effect of variation in CHI3L1 on serum YKL-40 ...	Yl
1	2008-06-16	18369459	Liu Y	2008-04-04	PLoS Genet	www.ncbi.nlm.nih.gov/pubmed/18369459	A genome-wide association study of psoriasis a...	
2	2008-06-16	18385676	Amos CI	2008-04-03	Nat Genet	www.ncbi.nlm.nih.gov/pubmed/18385676	Genome-wide association scan of tag SNPs ident...	I

3 rows × 34 columns

In [6]:
```
gwas[-3:] # Τρεις τελευταίες
```

Out[6]:

	DATE ADDED TO CATALOG	PUBMEDID	FIRST AUTHOR	DATE	JOURNAL	LINK	STUDY
64236	2018-01-12	29151059	Delgado DA	2017-11-18	J Med Genet	www.ncbi.nlm.nih.gov/pubmed/29151059	Genome-wide association study of telomere leng...
64237	2018-01-12	29151059	Delgado DA	2017-11-18	J Med Genet	www.ncbi.nlm.nih.gov/pubmed/29151059	Genome-wide association study of telomere leng...
64238	2018-01-12	29151059	Delgado DA	2017-11-18	J Med Genet	www.ncbi.nlm.nih.gov/pubmed/29151059	Genome-wide association study of telomere leng...

3 rows × 34 columns

In [7]:
```
gwas[["STUDY", "P-VALUE"]] # Μονο συγκεκριμμένες κολόνες
```

Out[7]:

	STUDY	P-VALUE
0	Effect of variation in CHI3L1 on serum YKL-40 ...	1e-13
1	A genome-wide association study of psoriasis a...	2e-06
2	Genome-wide association scan of tag SNPs ident...	3e-18
3	Genome-wide association scan of tag SNPs ident...	7e-06
4	Genome-wide association scan of tag SNPs ident...	8e-06
5	A susceptibility locus for lung cancer maps to...	5e-20
6	A variant associated with nicotine dependence,...	6e-20
7	Genome-wide association scan identifies a colo...	9e-26
8	Genome-wide association scan identifies a colo...	6e-10
9	Genome-wide association scan identifies a colo...	8e-28
10	A genome-wide association study identifies col...	3e-13
11	A genome-wide association study identifies col...	3e-18
12	Meta-analysis of genome-wide association data ...	5e-14
13	Meta-analysis of genome-wide association data ...	1e-10
14	Meta-analysis of genome-wide association data ...	1e-09
15	Meta-analysis of genome-wide association data ...	1e-09
16	Meta-analysis of genome-wide association data ...	1e-08
17	Genome-wide association study provides evidenc...	3e-08
18	A genome-wide association study in 574 schizop...	1e-06
19	SLC2A9 influences uric acid concentrations wit...	3e-70
20	SLC2A9 is a newly identified urate transporter...	3e-09
21	Newly identified genetic risk variants for cel...	3e-11
22	Newly identified genetic risk variants for cel...	4e-09
23	Newly identified genetic risk variants for cel...	1e-09
24	Newly identified genetic risk variants for cel...	5e-09
25	Newly identified genetic risk variants for cel...	7e-08
26	Multiple newly identified loci associated with...	9e-29
27	Multiple newly identified loci associated with...	2e-18
28	Multiple newly identified loci associated with...	2e-12
29	Multiple newly identified loci associated with...	6e-10
...	...	...
64209	Identification of a novel locus on chromosome ...	1e-09
64210	Identification of a novel locus on chromosome ...	4e-07
64211	Identification of a novel locus on chromosome ...	3e-07
64212	Genome-wide association study of pigmentary tr...	6e-06
64213	Genome-wide association study of pigmentary tr...	9e-06
64214	A GWAS Meta-analysis and Replication Study Ide...	6e-13
64215	A GWAS Meta-analysis and Replication Study Ide...	2e-12
64216	A GWAS Meta-analysis and Replication Study Ide...	5e-10
64217	A GWAS Meta-analysis and Replication Study Ide...	3e-08
64218	A GWAS Meta-analysis and Replication Study Ide...	1e-06

In [8]:
```
gwas[["STUDY", "P-VALUE"]][:3] # Sygkekrimmenes kolones, prwtes 3 grammes
```

Out[8]:

	STUDY	P-VALUE
0	Effect of variation in CHI3L1 on serum YKL-40 ...	1e-13
1	A genome-wide association study of psoriasis a...	2e-06
2	Genome-wide association scan of tag SNPs ident...	3e-18

Αυτό είναι ισοδύναμο με:

In [9]:
```
gwas[:3][["STUDY", "P-VALUE"]]
```

Out[9]:

	STUDY	P-VALUE
0	Effect of variation in CHI3L1 on serum YKL-40 ...	1e-13
1	A genome-wide association study of psoriasis a...	2e-06
2	Genome-wide association scan of tag SNPs ident...	3e-18

Λίστα με όλες τις κολόνες:

In [10]:
```
columns = list(gwas.columns.values)
columns
```

Out[10]:
```
['DATE ADDED TO CATALOG',
 'PUBMEDID',
 'FIRST AUTHOR',
 'DATE',
 'JOURNAL',
 'LINK',
 'STUDY',
 'DISEASE/TRAIT',
 'INITIAL SAMPLE SIZE',
 'REPLICATION SAMPLE SIZE',
 'REGION',
 'CHR_ID',
 'CHR_POS',
 'REPORTED GENE(S)',
 'MAPPED_GENE',
 'UPSTREAM_GENE_ID',
 'DOWNSTREAM_GENE_ID',
 'SNP_GENE_IDS',
 'UPSTREAM_GENE_DISTANCE',
 'DOWNSTREAM_GENE_DISTANCE',
 'STRONGEST SNP-RISK ALLELE',
 'SNPS',
 'MERGED',
 'SNP_ID_CURRENT',
 'CONTEXT',
 'INTERGENIC',
 'RISK ALLELE FREQUENCY',
 'P-VALUE',
 'PVALUE_MLOG',
 'P-VALUE (TEXT)',
 'OR or BETA',
 '95% CI (TEXT)',
 'PLATFORM [SNPS PASSING QC]',
 'CNV']
```

Όλες οι γραμμές που έχουν το γονίδιο BRCA2

In [11]:
```
gwas[gwas['MAPPED_GENE'] == 'BRCA2']
```

Out[11]:

	DATE ADDED TO CATALOG	PUBMEDID	FIRST AUTHOR	DATE	JOURNAL	LINK	STUD
9308	2013-09-12	23535733	Garcia-Closas M	2013-04-01	Nat Genet	www.ncbi.nlm.nih.gov/pubmed/23535733	Genome wid associatio studie identify fou .
10954	2014-05-12	24097068	Willer CJ	2013-10-06	Nat Genet	www.ncbi.nlm.nih.gov/pubmed/24097068	Discover an refinemer of loc associate wi.
17867	2013-09-12	23535729	Michailidou K	2013-04-01	Nat Genet	www.ncbi.nlm.nih.gov/pubmed/23535729	Large-scal genotypin identifies 4 new loci .
18831	2017-09-18	28604730	McKay JD	2017-06-12	Nat Genet	www.ncbi.nlm.nih.gov/pubmed/28604730	Large-scal associatio analysi identifie ne.
18934	2017-09-18	28604730	McKay JD	2017-06-12	Nat Genet	www.ncbi.nlm.nih.gov/pubmed/28604730	Large-scal associatio analysi identifie ne.
18988	2017-09-18	28604730	McKay JD	2017-06-12	Nat Genet	www.ncbi.nlm.nih.gov/pubmed/28604730	Large-scal associatio analysi identifie ne.
25094	2017-09-18	28604730	McKay JD	2017-06-12	Nat Genet	www.ncbi.nlm.nih.gov/pubmed/28604730	Large-scal associatio analysi identifie ne.
27113	2015-01-21	24880342	Wang Y	2014-06-01	Nat Genet	www.ncbi.nlm.nih.gov/pubmed/24880342	Rar variants c large effec in BRCA and CHE.
27115	2015-01-21	24880342	Wang Y	2014-06-01	Nat Genet	www.ncbi.nlm.nih.gov/pubmed/24880342	Rar variants c large effec in BRCA and CHE.
41192	2017-02-03	27197191	Fehringer G	2016-04-20	Cancer Res	www.ncbi.nlm.nih.gov/pubmed/27197191	Cross cance genome wid analysis c lung, ova.
41215	2017-02-03	27197191	Fehringer G	2016-04-20	Cancer Res	www.ncbi.nlm.nih.gov/pubmed/27197191	Cross cance genome wid analysis c lung, ova.
41316	2017-03-31	27117709	Couch FJ	2016-04-27	Nat Commun	www.ncbi.nlm.nih.gov/pubmed/27117709	Identificatio of four nove susceptibilit

Όλα τα διαφορετικά Diseases / Traits

In [12]:
```
gwas["DISEASE/TRAIT"].unique()
```

Out[12]:
```
array(['YKL-40 levels', 'Psoriasis', 'Lung cancer', ...,
       'Bipolar disorder lithium response (categorical) or schizophrenia',
       'Fractures (vertebral)', 'Response to mepolizumab in severe asthma'], d
type=object)
```

ή μπορούμε να πάρουμε μία λίστα:

In [13]:
```
list(gwas["DISEASE/TRAIT"].unique())
```

Out[13]:
```
['YKL-40 levels',
 'Psoriasis',
 'Lung cancer',
 'Nicotine dependence',
 'Colorectal cancer',
 'Type 2 diabetes',
 'Breast cancer',
 'Schizophrenia',
 'Urate levels',
 'Celiac disease',
 'Prostate cancer',
 'LDL cholesterol',
 'Fetal hemoglobin levels',
 'Recombination rate (females)',
 'Recombination rate (males)',
 'Iris color',
 'Systemic lupus erythematosus',
 'Type 1 diabetes',
 'HDL cholesterol',
 'Triglycerides',
 'Height',
 'Amyotrophic lateral sclerosis',
 'Coronary spasm',
 'Rheumatoid arthritis',
 'Blond vs. brown hair color',
 'Blue vs. green eyes',
 'Freckles',
 'Skin pigmentation',
 'Select biomarker traits',
 'Body mass index',
 'Waist circumference',
 'Sleep-related phenotypes',
 'Cystatin C',
 'Thyroid stimulating hormone',
 'Urinary albumin excretion',
 'Bone mineral density',
 'Hip geometry',
 'Atrial fibrillation',
 'Heart failure',
 'Major CVD',
 'Blood pressure',
 'Tonometry',
 'Morbidity-free survival',
 'Aging traits',
 'Diabetes related insulin traits',
 'Fasting plasma glucose',
 'Diabetes (incident)',
 'Electrocardiographic traits',
 'Heart rate variability traits',
 'Coronary artery calcification',
 'Subclinical atherosclerosis traits (other)',
 'Cognitive test performance',
 'Volumetric brain MRI',
 'Echocardiographic traits',
 'Endothelial function traits',
 'Exercise treadmill test traits',
 'Mean forced vital capacity from 2 exams',
 'Pulmonary function',
 'Factor VII',
 'Hemostatic factors and hematological phenotypes',
 "Crohn's disease",
 'F-cell distribution',
 'Glaucoma (exfoliation)',
 'Type 2 diabetes nephropathy',
 'Neuroticism',
 'Multiple sclerosis',
 'Asthma',
 'Obesity-related traits',
```

Όλες οι γραμμές που περιέχουν τον Brest στο Disease / Train

In [14]:
```
gwas[gwas["DISEASE/TRAIT"].str.contains("Breast")]
```

Out[14]:

	DATE ADDED TO CATALOG	PUBMEDID	FIRST AUTHOR	DATE	JOURNAL	LINK	STUDY
17	2008-06-16	18326623	Gold B	2008-03-11	Proc Natl Acad Sci U S A	www.ncbi.nlm.nih.gov/pubmed/18326623	Genome-wide association study provides evidenc..
126	2008-09-10	17903305	Murabito JM	2007-09-19	BMC Med Genet	www.ncbi.nlm.nih.gov/pubmed/17903305	A genome-wide association study o breast and ..
127	2008-09-10	17903305	Murabito JM	2007-09-19	BMC Med Genet	www.ncbi.nlm.nih.gov/pubmed/17903305	A genome-wide association study o breast and ..
128	2008-09-10	17903305	Murabito JM	2007-09-19	BMC Med Genet	www.ncbi.nlm.nih.gov/pubmed/17903305	A genome-wide association study o breast and ..
129	2008-09-10	17903305	Murabito JM	2007-09-19	BMC Med Genet	www.ncbi.nlm.nih.gov/pubmed/17903305	A genome-wide association study o breast and ..
130	2008-09-10	17903305	Murabito JM	2007-09-19	BMC Med Genet	www.ncbi.nlm.nih.gov/pubmed/17903305	A genome-wide association study o breast and ..
238	2008-06-16	17529967	Easton DF	2007-05-27	Nature	www.ncbi.nlm.nih.gov/pubmed/17529967	Genome-wide association study identifies novel..
239	2008-06-16	17529967	Easton DF	2007-05-27	Nature	www.ncbi.nlm.nih.gov/pubmed/17529967	Genome-wide association study identifies novel..
240	2008-06-16	17529967	Easton DF	2007-05-27	Nature	www.ncbi.nlm.nih.gov/pubmed/17529967	Genome-wide association study identifies novel..
241	2008-06-16	17529967	Easton DF	2007-05-27	Nature	www.ncbi.nlm.nih.gov/pubmed/17529967	Genome-wide association study identifies novel..
242	2008-06-16	17529967	Easton DF	2007-05-27	Nature	www.ncbi.nlm.nih.gov/pubmed/17529967	Genome-wide association study identifies novel..

Το ίδιο αλλά χωρίς να νοιαζόμαστε για μικρά/κεφαλαία (case insensitive search)

In [15]:
```
gwas[gwas["DISEASE/TRAIT"].str.contains("Breast", case=False)]
```

Out[15]:

	DATE ADDED TO CATALOG	PUBMEDID	FIRST AUTHOR	DATE	JOURNAL	LINK	STUDY
17	2008-06-16	18326623	Gold B	2008-03-11	Proc Natl Acad Sci U S A	www.ncbi.nlm.nih.gov/pubmed/18326623	Genome-wide association study provides evidenc..
126	2008-09-10	17903305	Murabito JM	2007-09-19	BMC Med Genet	www.ncbi.nlm.nih.gov/pubmed/17903305	A genome-wide association study o breast and ..
127	2008-09-10	17903305	Murabito JM	2007-09-19	BMC Med Genet	www.ncbi.nlm.nih.gov/pubmed/17903305	A genome-wide association study o breast and ..
128	2008-09-10	17903305	Murabito JM	2007-09-19	BMC Med Genet	www.ncbi.nlm.nih.gov/pubmed/17903305	A genome-wide association study o breast and ..
129	2008-09-10	17903305	Murabito JM	2007-09-19	BMC Med Genet	www.ncbi.nlm.nih.gov/pubmed/17903305	A genome-wide association study o breast and ..
130	2008-09-10	17903305	Murabito JM	2007-09-19	BMC Med Genet	www.ncbi.nlm.nih.gov/pubmed/17903305	A genome-wide association study o breast and ..
238	2008-06-16	17529967	Easton DF	2007-05-27	Nature	www.ncbi.nlm.nih.gov/pubmed/17529967	Genome-wide association study identifies novel..
239	2008-06-16	17529967	Easton DF	2007-05-27	Nature	www.ncbi.nlm.nih.gov/pubmed/17529967	Genome-wide association study identifies novel..
240	2008-06-16	17529967	Easton DF	2007-05-27	Nature	www.ncbi.nlm.nih.gov/pubmed/17529967	Genome-wide association study identifies novel..
241	2008-06-16	17529967	Easton DF	2007-05-27	Nature	www.ncbi.nlm.nih.gov/pubmed/17529967	Genome-wide association study identifies novel..
242	2008-06-16	17529967	Easton DF	2007-05-27	Nature	www.ncbi.nlm.nih.gov/pubmed/17529967	Genome-wide association study identifies novel..

Τύπωσε για όλα τα διαφορετικά γονίδια, πόσες γραμμές υπάρχουν

In [16]:
```
gwas["MAPPED_GENE"].value_counts()
```

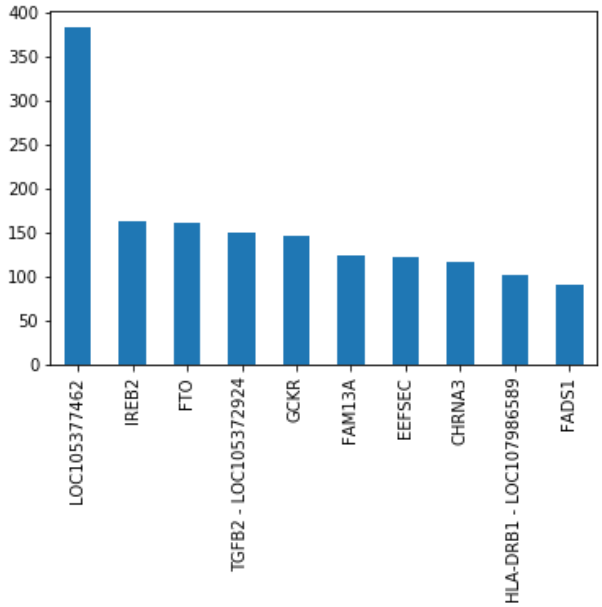
Out[16]:
```
LOC105377462                    384
IREB2                           163
FTO                             161
TGFB2 - LOC105372924            150
GCKR                            147
FAM13A                          124
EEFSEC                          123
CHRNA3                          116
HLA-DRB1 - LOC107986589         103
FADS1                            91
TCF7L2                           86
LOC107986647 - LOC105378010      85
TERT                             85
LOC105373352 - TMEM18            80
CHRNA5                           79
CDKN2B-AS1                       79
RUVBL1                           79
HYKK                             77
HLA-DQB1 - MTCO3P1               77
TRIB1 - LOC105375746             77
ABO - LCN1P2                     75
CSMD1                            73
LOC101928778 - SEC16B            70
HERPUD1 - CETP                   70
TMPRSS6                          68
ZPR1                             68
LOC105378797                     68
JAZF1                            67
VEGFA - LOC105375070             67
HLA-DRB9 - HLA-DRB5              66
                                ...
HIST2H3DP1 - RPL22P6              1
CUTC, COX15                       1
LOC107984948 - KLF17              1
ITPKB-IT1, ITPKB                  1
LOC105370174 - VWA8               1
CACNA1C, LOC107984540             1
RN7SKP279 - LOC107984373          1
AGO4                              1
C14orf159                         1
DNAL4                             1
MRPS27                            1
SCRN3                             1
ACD                               1
STRADA                            1
CAPZA2 - LOC105375465             1
OR52Q1P - LOC107984304            1
HKDC1, LOC101928994               1
LCE1F - LCE1E                     1
SUDS3                             1
LARS2, LARS2-AS1                  1
BAG3                              1
ZFP82                             1
NSMCE2                            1
LOC105376311                      1
MGST3 - ALDH9A1                   1
SLC1A1, SPATA6L                   1
MIR7515HG - LINC00487             1
LOC105373324                      1
APOOP3 - LEMD3                    1
PSMA3                             1
Name: MAPPED_GENE, Length: 19192, dtype: int64
```

Κάνε ένα bar plot για τα πρώτα 10 από αυτά:

In [17]:
```
import matplotlib.pyplot as plt
gwas["MAPPED_GENE"].value_counts()[:10].plot(kind="bar")
```

Out[17]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x114e77128>
```

In [18]:
```
plt.show()
```



Μπορούμε να διαγράψουμε ένα πεδίο:

In [19]:
```
gwas = gwas.drop('LINK', 1)
```

Μπορούμε να μετατρέψουμε ένα πεδίο σε ένα άλλο φορμάτ. Π.χ:

In [20]:
```
gwas['DATE'] = pd.to_datetime(gwas["DATE"]) # Μετατροπή του DATE από string σε
datetime
```

Τώρα μπορούμε να κάνουμε sort τα δεδομένα μας με βάση το DATE:

In [21]:
```
gwas_date_sorted = gwas.sort_values('DATE')
```

Και μπορούμε να κάνουμε plot όλα τα p-values με βάση το DATE που έγιναν publish

In [22]:
```
gwas_date_sorted['P-VALUE'].plot()
```

```
---------------------------------------------------------------------------
TypeError                                 Traceback (most recent call last)
<ipython-input-22-d0f6d99f691a> in <module>()
----> 1 gwas_date_sorted['P-VALUE'].plot()

~/anaconda3/envs/arkalos/lib/python3.6/site-packages/pandas/plotting/_core.py
in __call__(self, kind, ax, figsize, use_index, title, grid, legend, style, lo
gx, logy, loglog, xticks, yticks, xlim, ylim, rot, fontsize, colormap, table,
yerr, xerr, label, secondary_y, **kwds)
   2451                            colormap=colormap, table=table, yerr=yerr,
   2452                            xerr=xerr, label=label, secondary_y=seconda
ry_y,
-> 2453                            **kwds)
   2454     __call__.__doc__ = plot_series.__doc__
   2455

~/anaconda3/envs/arkalos/lib/python3.6/site-packages/pandas/plotting/_core.py
in plot_series(data, kind, ax, figsize, use_index, title, grid, legend, style,
logx, logy, loglog, xticks, yticks, xlim, ylim, rot, fontsize, colormap, tabl
e, yerr, xerr, label, secondary_y, **kwds)
   1892                  yerr=yerr, xerr=xerr,
   1893                  label=label, secondary_y=secondary_y,
-> 1894                  **kwds)
   1895
   1896

~/anaconda3/envs/arkalos/lib/python3.6/site-packages/pandas/plotting/_core.py
in _plot(data, x, y, subplots, ax, kind, **kwds)
   1692         plot_obj = klass(data, subplots=subplots, ax=ax, kind=kind, **
kwds)
   1693
-> 1694     plot_obj.generate()
   1695     plot_obj.draw()
   1696     return plot_obj.result

~/anaconda3/envs/arkalos/lib/python3.6/site-packages/pandas/plotting/_core.py
in generate(self)
    241     def generate(self):
    242         self._args_adjust()
--> 243         self._compute_plot_data()
    244         self._setup_subplots()
    245         self._make_plot()

~/anaconda3/envs/arkalos/lib/python3.6/site-packages/pandas/plotting/_core.py
in _compute_plot_data(self)
    350         if is_empty:
    351             raise TypeError('Empty {0!r}: no numeric data to '
--> 352                             'plot'.format(numeric_data.__class__.__nam
e__))
    353
    354         self.data = numeric_data

TypeError: Empty 'DataFrame': no numeric data to plot
```

Τι έγινε εδώ;

In [23]:
```
set([type(x) for x in gwas_date_sorted['P-VALUE']])
```

Out[23]:
```
{float, str}
```

Κάποια p-values είναι str και κάποια float! Ποια είναι strings? Ας προσπαθήσουμε να τα μετατρέψουμε όλα σε numberic:

In [24]:
```
pd.to_numeric(gwas_date_sorted['P-VALUE'])
```

```
---------------------------------------------------------------------------
ValueError                                Traceback (most recent call last)
pandas/_libs/src/inference.pyx in pandas._libs.lib.maybe_convert_numeric (pand
as/_libs/lib.c:56156)()

ValueError: Unable to parse string "2E-1449"

During handling of the above exception, another exception occurred:

ValueError                                Traceback (most recent call last)
<ipython-input-24-8a7ceacdb8bc> in <module>()
----> 1 pd.to_numeric(gwas_date_sorted['P-VALUE'])

~/anaconda3/envs/arkalos/lib/python3.6/site-packages/pandas/core/tools/numeri
c.py in to_numeric(arg, errors, downcast)
    124             coerce_numeric = False if errors in ('ignore', 'raise') el
se True
    125             values = lib.maybe_convert_numeric(values, set(),
--> 126                                                coerce_numeric=coerce_n
umeric)
    127
    128     except Exception:

pandas/_libs/src/inference.pyx in pandas._libs.lib.maybe_convert_numeric (pand
as/_libs/lib.c:56638)()

ValueError: Unable to parse string "2E-1449" at position 35122
```

Αποτυχία. Ας του πούμε να βάλει NaN values όπου η μετατροπή αποτυγχάνει:

In [25]:
```
pd.to_numeric(gwas_date_sorted['P-VALUE'], errors='coerce')
```

Out[25]:
```
274      4.000000e-08
273      8.000000e-06
272      1.000000e-10
271      2.000000e-06
270      2.000000e-06
269      7.000000e-06
268      8.000000e-12
408      5.000000e-10
266      2.000000e-06
267      6.000000e-06
336      6.000000e-08
409      2.000000e-34
265      3.000000e-06
410      9.000000e-06
264      6.000000e-06
262      3.000000e-06
261      2.000000e-06
260      7.000000e-07
263      5.000000e-06
411      2.000000e-18
259      2.000000e-12
747      1.000000e-06
258      4.000000e-07
41851    3.000000e-15
256      1.000000e-12
257      9.000000e-13
414      1.000000e-39
413      6.000000e-18
412      2.000000e-14
324      2.000000e-06
              ...
64151    7.000000e-09
64152    2.000000e-08
64153    4.000000e-11
64154    3.000000e-09
64155    3.000000e-13
64148    7.000000e-13
64138    3.000000e-08
64139    2.000000e-09
64136    3.000000e-10
64120    9.000000e-09
64121    1.000000e-09
64122    9.000000e-11
64078    7.000000e-07
64124    1.000000e-08
64125    6.000000e-11
64137    1.000000e-12
64127    3.000000e-09
64126    1.000000e-08
64129    4.000000e-13
64130    4.000000e-09
64131    1.000000e-09
64132    3.000000e-09
64133    4.000000e-08
64134    5.000000e-12
64135    3.000000e-08
64128    2.000000e-16
64123    9.000000e-11
64032    8.000000e-08
64031    3.000000e-07
64234    4.000000e-06
Name: P-VALUE, Length: 64239, dtype: float64
```

Ωραία ας αντικαταστήσουμε τώρα το παλιό με το νέο πεδίο:

In [26]:
```
gwas_date_sorted['P-VALUE'] = pd.to_numeric(gwas_date_sorted['P-VALUE'], error
s='coerce')
```

Τώρα μπορούμε να δούμε ποιες γραμμή είναι NaN

In [27]:
```
gwas_date_sorted[gwas_date_sorted['P-VALUE'].isnull()]
```

Out[27]:

	DATE ADDED TO CATALOG	PUBMEDID	FIRST AUTHOR	DATE	JOURNAL	STUDY	DISEASE/TRAIT	INITIAL SAMPLE SIZE	F S
37312	2016-12-01	26910538	Choi SH	2016-02-24	PLoS Genet	Six Novel Loci Associated with Circulating VEG...	Vascular endothelial growth factor levels	9541 European ancestry individuals, 1,115 Cile...	

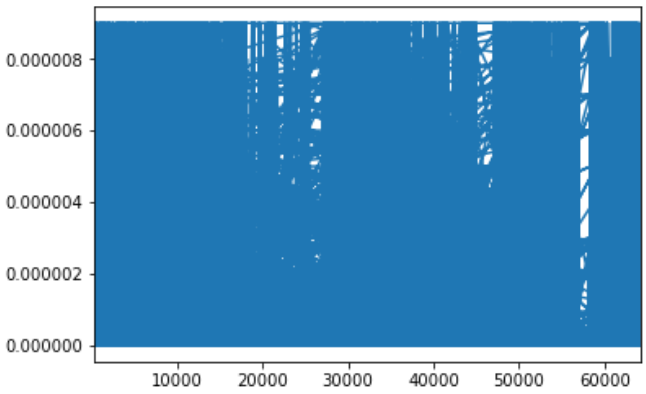
1 rows × 33 columns

Μία γραμμή δημιουργούσε αυτό το πρόβλημα. Ας τη βγάλουμε:

In [28]:
```
gwas_date_sorted = gwas_date_sorted[~gwas_date_sorted['P-VALUE'].isnull()]
```

Τώρα μπορούμε να κάνουμε το plot:

In [29]:
```
gwas_date_sorted['P-VALUE'].plot()
plt.show()
```



Τι είναι αυτό καλέ; Από default στον X άξονα βάζει το index του dataframe:

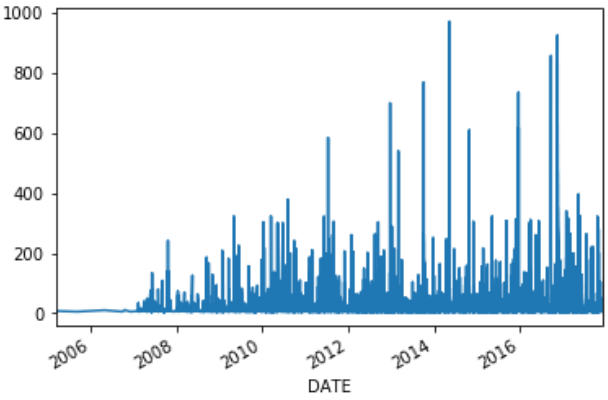
In [30]:
```
gwas_date_sorted.index
```

Out[30]:
```
Int64Index([  274,   273,   272,   271,   270,   269,   268,   408,   266,
              267,
             ...
             64131, 64132, 64133, 64134, 64135, 64128, 64123, 64032, 64031,
             64234],
            dtype='int64', length=64238)
```

Το index είναι ένα μοναδικό στοιχείο που χαρακτηρίζει κάθε γραμμή. Από default περιέχει τον άυξων αριθμό της γραμμής στο CSV αρχείο. Μπορούμε όμως να αλλάξουμε το index:

In [31]:
```
gwas_date_sorted2 = gwas_date_sorted.set_index(gwas_date_sorted['DATE'])
```

In [32]:
```
gwas_date_sorted2['PVALUE_MLOG'].plot()
plt.show()
```



Παρατηρούμε ότι όσο περνάει ο χρόνος. Οι GWAS έρευνες που γίνονται έχουν πιο χαμηλό p-value.

Μπορούμε να φτιάξουμε ένα νέα field μέσω του index

In [33]:
```
gwas_date_sorted2['YEAR'] = gwas_date_sorted2.index.year
```

Μπορούμε επίσης να "γκρουπάρουμε" όλες τις γραμμές ανάλογα με τις τιμές ενός πεδίου:

In [34]:
```
gwas_date_sorted2.groupby('YEAR').aggregate('count')
```

Out[34]:

YEAR	DATE ADDED TO CATALOG	PUBMEDID	FIRST AUTHOR	DATE	JOURNAL	STUDY	DISEASE/TRAIT	INITIAL SAMPLE SIZE	REPLICATIO SAMPLE SIZ
2005	2	2	2	2	2	2	2	2	
2006	8	8	8	8	8	8	8	8	
2007	440	440	440	440	440	440	440	440	19
2008	977	977	977	977	977	977	977	977	62
2009	1390	1390	1390	1390	1390	1390	1390	1390	83
2010	2582	2582	2582	2582	2582	2582	2582	2582	125
2011	2602	2602	2602	2602	2602	2602	2602	2602	147
2012	4404	4404	4404	4404	4404	4404	4404	4404	160
2013	5551	5551	5551	5551	5551	5551	5551	5551	227
2014	4337	4337	4337	4337	4337	4337	4337	4337	187
2015	11972	11972	11972	11972	11972	11972	11972	11972	387
2016	14806	14806	14806	14806	14806	14806	14806	14806	232
2017	15167	15167	15167	15167	15167	15167	15167	15167	801

13 rows × 33 columns

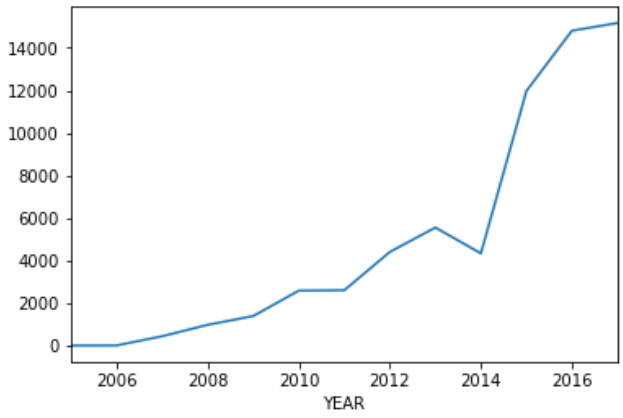
Το aggregate εφαρμόζει μία συνάρτηση σε κάθε ένα group ξεχωριστά. Υπάρχουν πολλές built-in συναρτήσεις όπως οι count, mean, median, sum, min, max.

Π.χ: πλοτάρουμε το πλήθος των entries ανά χρόνο

In [35]:
```
gwas_date_sorted2.groupby('YEAR').aggregate('count')['JOURNAL'].plot()
```

Out[35]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x12077fa90>
```

In [36]:
```
plt.show()
```



Τυπώνουμε όλες τις γραμμές που έχουν MAPPED_GENE το BRCA2 και έχουν p-value<0.0000001

In [37]:
```
gwas_date_sorted2[(gwas_date_sorted2["MAPPED_GENE"] == "BRCA2") & (gwas_date_so
rted2['P-VALUE']< 0.000000001)]
```

Out[37]:

DATE	DATE ADDED TO CATALOG	PUBMEDID	FIRST AUTHOR	DATE	JOURNAL	STUDY	DISEASE/TRAIT	SAI
2013-10-06	2014-05-12	24097068	Willer CJ	2013-10-06	Nat Genet	Discovery and refinement of loci associated wi...	LDL cholesterol	
2014-06-01	2015-01-21	24880342	Wang Y	2014-06-01	Nat Genet	Rare variants of large effect in BRCA2 and CHE...	Lung cancer	3,44 aden
2014-06-01	2015-01-21	24880342	Wang Y	2014-06-01	Nat Genet	Rare variants of large effect in BRCA2 and CHE...	Lung cancer	3,44 aden
2016-04-20	2017-02-03	27197191	Fehringer G	2016-04-20	Cancer Res	Cross-cancer genome-wide analysis of lung, ova...	Cancer	5,02 ar ca
2016-04-20	2017-02-03	27197191	Fehringer G	2016-04-20	Cancer Res	Cross-cancer genome-wide analysis of lung, ova...	Cancer (pleiotropy)	5,02 ar ca
2017-02-21	2017-06-26	28334899	Spracklen CN	2017-02-21	Hum Mol Genet	Association analyses of East Asian individuals...	LDL cholesterol levels	Asi
2017-06-12	2017-09-18	28604730	McKay JD	2017-06-12	Nat Genet	Large-scale association analysis identifies ne...	Lung cancer	anc
2017-06-12	2017-09-18	28604730	McKay JD	2017-06-12	Nat Genet	Large-scale association analysis identifies ne...	Squamous cell lung carcinoma	7,42 anc
2017-10-23	2017-12-19	29058716	Milne RL	2017-10-23	Nat Genet	Identification of ten variants associated with...	Breast cancer (estrogen-receptor negative)	anc
2017-10-23	2017-11-30	29059683	Michailidou K	2017-10-23	Nature	Association analysis identifies 65 new breast ...	Breast cancer	anc

10 rows × 34 columns

In [38]:
```
g=gwas_date_sorted2
```

Όλες οι γραμμές που έχουν χρωμόσωμα που ανοίκει στον πίνακα ['1', '2', ... '22', 'X','Y']

In [39]:
```
g3 = g[g["CHR_ID"].isin([str(x) for x in range(1,23)] + ["X", "Y"])]
```

Πόσες γραμμές ανά χρωμόσωμα έχουμε;

In [40]:
```
g3["CHR_ID"].value_counts()
```

Out[40]:
```
6     6067
1     5529
2     5039
3     4122
4     3554
5     3485
11    3348
7     3035
12    2912
8     2791
10    2755
15    2534
9     2512
16    2318
19    2177
17    2130
14    1640
20    1440
13    1405
18    1278
22    1091
21     550
X      372
Y        2
Name: CHR_ID, dtype: int64
```

Ποιο είναι το πιο χαμηλό p-value ανά χρωμόσωμα;

In [41]:
```
g3.groupby("CHR_ID")["P-VALUE"].aggregate('min')
```

Out[41]:
```
CHR_ID
1       0.000000e+00
10      0.000000e+00
11      0.000000e+00
12      0.000000e+00
13     9.000000e-256
14     2.000000e-188
15     1.000000e-300
16      0.000000e+00
17     2.000000e-298
18     3.000000e-200
19      0.000000e+00
2       0.000000e+00
20     2.000000e-200
21     4.000000e-104
22     5.000000e-178
3       0.000000e+00
4       0.000000e+00
5      5.000000e-274
6       0.000000e+00
7       0.000000e+00
8      5.000000e-217
9      1.000000e-312
X      1.000000e-247
Y       9.000000e-07
Name: P-VALUE, dtype: float64
```

Μπορούμε να σώσουμε ένα pandas αντικείμενο σε csv (ή κάποιο άλλο φορμάτ):

In [42]:
```
g3.to_csv('results.csv')
```

Και σε excel. Για να γίνει αυτό χρειάζεται το πακέτο xlwt:

In [44]:
```
!pip install xlwt
```

```
Collecting xlwt
  Downloading xlwt-1.3.0-py2.py3-none-any.whl (99kB)
    100% |████████████████████████████████| 102kB 546kB/s a 0:00:01
Installing collected packages: xlwt
Successfully installed xlwt-1.3.0
```

In [45]:
```
g3.to_excel('results.xls')
```

Ένα άλλος παράδειγμα: Από όλα τα studies που έχουν το Breast στο DISEASE/TRAIN και έχουν PVALUE<10^-10, βρες το χρωμόσωμα που έχει τα περισσότερα studies

In [46]:
```
g3[ (g3["DISEASE/TRAIT"].str.contains('Breast')) & (g3["PVALUE_MLOG"]>10)].grou
pby("CHR_ID")['JOURNAL'].aggregate('count').idxmax()
```

Out[46]:
```
'5'
```

Αναμενόμενο αφού το BRCA2 βρίσκεται στο χρωμόσωμα 5

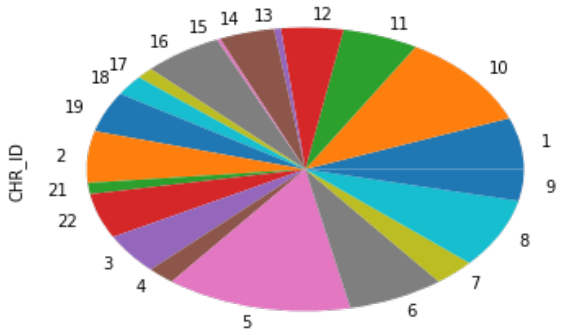
Να και ένα pie-chart με την κατανομή των studies ανά χρωμόσωμα:

ΠΡΟΣΟΧΗ! Ποτέ μην χρησιμοποιείται pie-chart σε (σοβαρές) ανοφορές/paper. Διαβάστε αυτό (http://www.businessinsider.com/pie-charts-are-the-worst-2013-6) και αυτό (https://blog.funnel.io/why-we-dont-use-pie-charts-and-some-tips-on-better-data-visualizations).

In [47]:
```
g3[ (g3["DISEASE/TRAIT"].str.contains('Breast')) & (g3["PVALUE_MLOG"]>10)].grou
pby("CHR_ID")['CHR_ID'].aggregate('count').plot(kind='pie')
```

Out[47]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x12b0d2550>
```

In [48]:
```
plt.show()
```



Μερικά ακόμα παραδείγματα:

Ποιος είναι ο ερευνητής που έχει τις περισσότερες δημοσιεύσεις στο Nature Genetics;

In [50]:
```
g3[g3['JOURNAL'] == 'Nat Genet'].groupby(
    'FIRST AUTHOR').aggregate('count')['PUBMEDID'].idxmax()
```

Out[50]:
```
'Pickrell JK'
```

Ποιο region περιέχει τις περισσότερες μελέτες σχετικά με καρκίνο;

In [51]:
```
g3[g3['DISEASE/TRAIT'].str.contains('cancer', case=False)].groupby(
    'REGION').aggregate('count')['PUBMEDID'].idxmax()
```

Out[51]:
```
'8q24.21'
```

Ποιος είναι ο μέσος όρος και το median του allele_frequency για όλα τα variants που ανακαλύπτοντε κάθε χρόνο;

In [53]:
```
# Μετατρέπουμε το ALLELE FREQUENCY σε numeric (από string)
g3['RISK ALLELE FREQUENCY'] = pd.to_numeric(g3['RISK ALLELE FREQUENCY'], error
s='coerce')

# Ο Μέσος όρος. ΠΡΟΣΟΧΗ! Αφαιρούμε όσα έχουν null RISK ALLELE FREQUENCY
g3[~g3['RISK ALLELE FREQUENCY'].isnull()].groupby('YEAR')['RISK ALLELE FREQUENC
Y'].aggregate('mean')
```

```
/Users/alexandroskanterakis/anaconda3/envs/arkalos/lib/python3.6/site-packages
/ipykernel_launcher.py:2: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: http://pandas.pydata.org/pandas-docs/sta
ble/indexing.html#indexing-view-versus-copy
```

Out[53]:
```
YEAR
2006    0.370000
2007    0.415852
2008    0.391309
2009    0.364374
2010    0.359003
2011    0.389704
2012    0.332810
2013    0.413206
2014    0.425324
2015    0.508952
2016    0.366965
2017    0.387201
Name: RISK ALLELE FREQUENCY, dtype: float64
```

In [54]:
```
# Το median
g3[~g3['RISK ALLELE FREQUENCY'].isnull()].groupby('YEAR')['RISK ALLELE FREQUENC
Y'].aggregate('median')
```

Out[54]:
```
YEAR
2006    0.370000
2007    0.400000
2008    0.350000
2009    0.320000
2010    0.330000
2011    0.350000
2012    0.290000
2013    0.378664
2014    0.400000
2015    0.494200
2016    0.336300
2017    0.359500
Name: RISK ALLELE FREQUENCY, dtype: float64
```

Και τα δυο μαζί:

In [81]:
```
g4 = g3[~g3['RISK ALLELE FREQUENCY'].isnull()].groupby('YEAR')['RISK ALLELE FREQUENCY'].aggregate(['mean', 'median'])
g4
```

Out[81]:

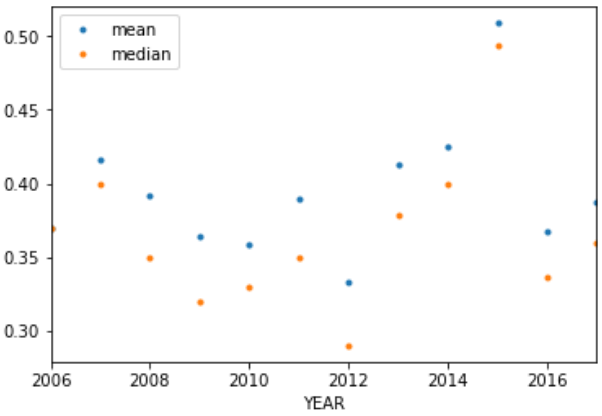
YEAR	mean	median
2006	0.370000	0.370000
2007	0.415852	0.400000
2008	0.391309	0.350000
2009	0.364374	0.320000
2010	0.359003	0.330000
2011	0.389704	0.350000
2012	0.332810	0.290000
2013	0.413206	0.378664
2014	0.425324	0.400000
2015	0.508952	0.494200
2016	0.366965	0.336300
2017	0.387201	0.359500

Ας κάνουμε ένα scatter plot με τον x να είναι το YEAR και το y να είναι τα mean και median

In [71]:
```
g4.plot(style='.')
```

Out[71]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x12077ff60>
```

In [72]:
```
plt.show()
```

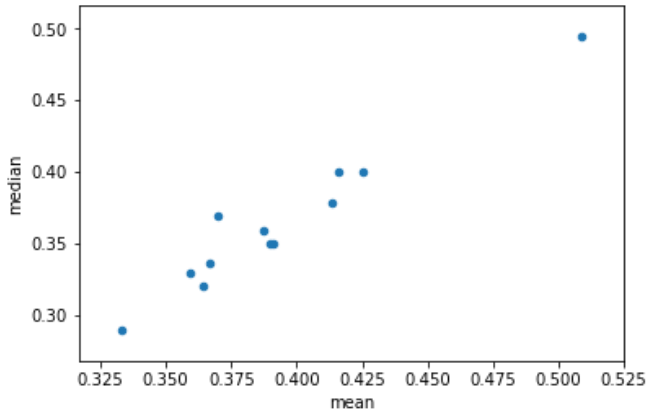


Και ένα scatter plot με το χ να είναι το mean και το y το median:

In [78]:
```
g4.plot.scatter(x='mean', y='median')
```

Out[78]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x129569860>
```

In [79]:
```
plt.show()
```



Ένας (από τους πολλούς) τρόπους για να φτιάξετε ένα data frame από δικά σας δεδομένα είναι:

In [31]:
```
data={"A": [1,2,3,4], "B": ["aa", "bb", "cc", "dd"], "C": [True, False, False, True]}
df = pd.DataFrame(data)
df
```

Out[31]:

	A	B	C
0	1	aa	True
1	2	bb	False
2	3	cc	False
3	4	dd	True

Πως επιλέγουμε συγκεκριμένες γραμμές

In [7]:
```
df.iloc[:2]
```

Out[7]:

	A	B	C
0	1	aa	True
1	2	bb	False

In [8]:
```
df[:2]
```

Out[8]:

	A	B	C
0	1	aa	True
1	2	bb	False

Πάρε μόνο τις γραμμές 2 και 4

In [13]:
```
df.loc[[1,3]]
```

Out[13]:

	A	B	C
1	2	bb	False
3	4	dd	True

Μετατροπη του DataFrame σε dictionary:

In [16]:
```
df.to_dict('index') # Αυτό φαίνεται πιο.. κατανοητό..
```

Out[16]:
```
{0: {'A': 1, 'B': 'aa', 'C': True},
 1: {'A': 2, 'B': 'bb', 'C': False},
 2: {'A': 3, 'B': 'cc', 'C': False},
 3: {'A': 4, 'B': 'dd', 'C': True}}
```

In [15]:
```
df.to_dict()
```

Out[15]:
```
{'A': {0: 1, 1: 2, 2: 3, 3: 4},
 'B': {0: 'aa', 1: 'bb', 2: 'cc', 3: 'dd'},
 'C': {0: True, 1: False, 2: False, 3: True}}
```

Προσθήκη νέας κολόνας από συνδοιασμό άλλω κολόνων:

In [32]:
```
df['D'] = df['A'].apply(lambda x:x**2)
df
```

Out[32]:

	A	B	C	D
0	1	aa	True	1
1	2	bb	False	4
2	3	cc	False	9
3	4	dd	True	16

In [33]:
```
df['E'] = df['C'].map({True: 'Male', False: 'Female'})
df
```

Out[33]:

	A	B	C	D	E
0	1	aa	True	1	Male
1	2	bb	False	4	Female
2	3	cc	False	9	Female
3	4	dd	True	16	Male

Αλλαγή του ονόματος των στηλών:

In [34]:
```
df = df.rename(index=str, columns={'C': 'Is Male', 'B': 'City'})
df
```

Out[34]:

	A	City	Is Male	D	E
0	1	aa	True	1	Male
1	2	bb	False	4	Female
2	3	cc	False	9	Female
3	4	dd	True	16	Male

Αλλαγή του ονόματων του index

In [35]:
```
df.index = df.index.rename('Εγγραφές')
df
```

Out[35]:

	A	City	Is Male	D	E
Εγγραφές					
0	1	aa	True	1	Male
1	2	bb	False	4	Female
2	3	cc	False	9	Female
3	4	dd	True	16	Male