Data_cleaning_challeng

July 17, 2018

1 Challenge: Data cleaning & validation

```
In [71]: # Import all libraries needed
         import pandas as pd
         import numpy as np
         import re
```

1.1 Importing raw data

```
In [72]: # bringing in raw data, encoding is latin_1
         raw = pd.read_csv('WELLCOME_APCspend2013_forThinkful.csv', encoding='latin_1')
         raw.head(5)
```

```
Out[72]:               PMID/PMCID Publisher            Journal title  \
         0                    NaN       CUP   Psychological Medicine
         1             PMC3679557       ACS        Biomacromolecules
         2  23043264  PMC3506128        ACS               J Med Chem
         3    23438330 PMC3646402       ACS               J Med Chem
         4   23438216 PMC3601604        ACS               J Org Chem

                                                 Article title  \
         0  Reduced parahippocampal cortical thickness in ...
         1  Structural characterization of a Model Gram-ne...
         2  Fumaroylamino-4,5-epoxymorphinans and related ...
         3  Orvinols with mixed kappa/mu opioid receptor a...
         4  Regioselective opening of myo-inositol orthoes...

           COST (č) charged to Wellcome (inc VAT when charged)
         0                                              č0.00
         1                                           č2381.04
         2                                            č642.56
         3                                            č669.64
         4                                            č685.88
```

1.2 Look at data

```
In [73]: cleaning_df = raw
```

```
         cleaning_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2127 entries, 0 to 2126
Data columns (total 5 columns):
PMID/PMCID                                             1928 non-null object
Publisher                                              2127 non-null object
Journal title                                          2126 non-null object
Article title                                          2127 non-null object
COST (č) charged to Wellcome (inc VAT when charged)    2127 non-null object
dtypes: object(5)
memory usage: 83.2+ KB
```

PMID and Journal title have null values

1.2.1 Cleaning PMID/PMCID column

```
In [74]: #cleaning_df['PMID/PMCID'].value_counts()
         #commented out becuase not important
```

There is plenty of work here, but we can definitely see that not all columns have an ID number, so we can start cleaning those.

```
In [75]: # Cleaning ovious ones that don't have a number by putting None in the field
         # cleaning_df['PMID/PMCID'] = cleaning_df['PMID/PMCID'].str.replace('-', 'None')
         # cleaning_df['PMID/PMCID'] = cleaning_df['PMID/PMCID'].str.replace('Not yet available'
         # cleaning_df['PMID/PMCID'] = cleaning_df['PMID/PMCID'].str.replace('In Process', 'None
         # cleaning_df['PMID/PMCID'] = cleaning_df['PMID/PMCID'].str.replace('print in press', '

         #commented out because not using this column
```

```
In [76]: cleaning_df['Journal title'].value_counts()
```

```
Out[76]: PLoS One
         PLoS ONE
         Journal of Biological Chemistry
         Nucleic Acids Research
         Proceedings of the National Academy of Sciences
         PLoS Neglected Tropical Diseases
         Human Molecular Genetics
         Nature Communications
         Neuroimage
         PLoS Pathogens
         PLoS Genetics
         Brain
         BMC Public Health
         NeuroImage
         PLOS ONE
```

```
Movement Disorders
Biochemical Journal
Developmental Cell
Journal of Neuroscience
Journal of General Virology
PLOS One
BMJ
Current Biology
Neuron
Cell Reports
Journal of Virology
Journal of Structural Biology
Journal of Cell Science
Molecular Microbiology
Journal of Physiology

N Biotechnol.
PLOS Computational Biology
Neuroimage: clinical
Journal of The American Society for Mass Spectrometry
Journal of Nutrition Education and Behaviour
Trop Med Int Health
Cost Effectiveness and Resource Allocation
BMC  Public Health
Antioxidants & Redox Signaling
Journal of Applied Crystallography
Hernia
Developmental Science
Database
Health: An Interdisciplinary Journal for the Social Study of Health, Illness and Medici
Current Opinion in Neurobiology
BBA - Molecular Basis of Disease
American Journal of Psychiatry
Statistics in Medicine
Molecular & Cellular Proteomics
BMC Infectious Diseases
Journal of Cultural Economy
Journal of Psychopharmacology
Scientific Reports-11-00861B
RESPIRATORY RESEARCH
Organic & Biomolecular Chemistry
Journal of Virol
STRUCTURE
Public Library of Science ONE
Neuropharmacology
BMC Neurology
Name: Journal title, Length: 984, dtype: int64
```

There are some journals that can be cleaned up. E.g., Acta Crystallographica is a mess

```
In [77]: cleaning_df['journal_lc']= cleaning_df['Journal title'].str.lower()
         cleaning_df.head()
```

```
Out[77]:               PMID/PMCID Publisher          Journal title  \
         0                    NaN       CUP  Psychological Medicine
         1             PMC3679557       ACS       Biomacromolecules
         2  23043264  PMC3506128       ACS                J Med Chem
         3    23438330 PMC3646402       ACS                J Med Chem
         4   23438216 PMC3601604       ACS                J Org Chem

                                                Article title  \
         0  Reduced parahippocampal cortical thickness in ...
         1  Structural characterization of a Model Gram-ne...
         2  Fumaroylamino-4,5-epoxymorphinans and related ...
         3  Orvinols with mixed kappa/mu opioid receptor a...
         4  Regioselective opening of myo-inositol orthoes...

           COST (č) charged to Wellcome (inc VAT when charged)              journal_lc
         0                                              č0.00  psychological medicine
         1                                           č2381.04       biomacromolecules
         2                                            č642.56              j med chem
         3                                            č669.64              j med chem
         4                                            č685.88              j org chem
```

```
In [78]: cleaning_df['journal_lc'].value_counts()
```

```
Out[78]: plos one                                                               190
         journal of biological chemistry                                         53
         neuroimage                                                              29
         plos pathogens                                                          24
         plos genetics                                                           24
         nucleic acids research                                                  23
         plos neglected tropical diseases                                        20
         proceedings of the national academy of sciences                         20
         nature communications                                                   19
         human molecular genetics                                                19
         brain                                                                   14
         bmc public health                                                       14
         movement disorders                                                      13
         biochemical journal                                                     12
         developmental cell                                                      12
         journal of neuroscience                                                 12
         journal of general virology                                             11
         current biology                                                         11
         bmj                                                                     10
         bmj open                                                                 9
         cell reports                                                             9
         neuron                                                                9
         plosone                                                               9
         plos computational biology                                            9
         journal of cell science                                               8
         hepatology                                                            8
         journal of physiology                                                 8
         proceedings of the royal society b: biological sciences               8
         journal of structural biology                                         8
         malaria journal                                                       8
                                                                            ...
         aging cell                                                            1
         molecular genetics and metabolism                                     1
         pflugers archiv                                                       1
         trends in microbiology                                                1
         theranostics                                                          1
         current obstetrics and gynecology reports                             1
         neurogenetics                                                         1
         journal of cheminformatics                                            1
         parasit vectors.                                                      1
         clinical transcriptional science                                      1
         developmental biology                                                 1
         medical humanities                                                    1
         chest                                                                 1
         social science & medicine                                             1
         journal of mechanisms of ageing and development                       1
         angewandte chemie                                                     1
         international reviews of immunology                                   1
         frontiers in developmental psychology                                 1
         mol biol and evolution                                                1
         journal of medicinal chemistry                                        1
         j biol chemistry                                                      1
         international journal of health geographics                           1
         neuropharmacology                                                     1
         cell press - cell reports                                             1
         frontiers in immunology                                               1
         plos  computational biology                                           1
         cerebral cortex print                                                 1
         genesis: journal of genetics                                          1
         birth defects research part a: clinical and molecular teratology      1
         journal of computational neuroscience                                 1
         Name: journal_lc, Length: 928, dtype: int64
```

```
In [79]: cleaning_df[cleaning_df['journal_lc'].isna()]
```

```
Out[79]:     PMID/PMCID  Publisher Journal title  \
         986        NaN  MacMillan           NaN

                                                  Article title  \
         986  Fungal Disease in Britain and the United State...

             COST (č) charged to Wellcome (inc VAT when charged) journal_lc
         986                                          č13200.00        NaN
```

```
In [80]: cleaning_df.drop(index=986, inplace=True)
```

```
In [81]: cleaning_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 2126 entries, 0 to 2126
Data columns (total 6 columns):
PMID/PMCID                                             1928 non-null object
Publisher                                              2126 non-null object
Journal title                                          2126 non-null object
Article title                                          2126 non-null object
COST (č) charged to Wellcome (inc VAT when charged)    2126 non-null object
journal_lc                                             2126 non-null object
dtypes: object(6)
memory usage: 116.3+ KB
```

```
In [82]: cleaning_df[cleaning_df['journal_lc'].str.contains('plos')]['journal_lc'].value_counts(
         # cleaning_df['journal_lc'].value_counts()
```

```
Out[82]: plos one                              190
         plos pathogens                         24
         plos genetics                          24
         plos neglected tropical diseases       20
         plos computational biology              9
         plosone                                 9
         plos 1                                  7
         plos                                    4
         plos medicine                           4
         plos biology                            2
         plos  computational biology             1
         plos  one                               1
         plos medicine journal                   1
         plos negected tropical diseases         1
         plos ntd                                1
         Name: journal_lc, dtype: int64
```

1.2.2 Removing spaces

```
In [83]: cleaning_df['journal_lc'] = cleaning_df['journal_lc'].str.replace('  ', ' ')
         cleaning_df['journal_lc'] = cleaning_df['journal_lc'].str.strip()
```

```
In [84]: cleaning_df[cleaning_df['journal_lc'].str.contains('plos')]['journal_lc'].value_counts(
Out[84]: plos one                            191
         plos pathogens                       24
         plos genetics                        24
         plos neglected tropical diseases     20
         plos computational biology           10
         plosone                               9
         plos 1                                7
         plos                                  4
         plos medicine                         4
         plos biology                          2
         plos medicine journal                 1
         plos negected tropical diseases       1
         plos ntd                              1
         Name: journal_lc, dtype: int64
```

```
In [85]: cleaning_df['journal_lc'].value_counts()
```

```
Out[85]: plos one                                                   191
         journal of biological chemistry                             53
         neuroimage                                                  29
         nucleic acids research                                      26
         plos genetics                                               24
         plos pathogens                                              24
         proceedings of the national academy of sciences             22
         plos neglected tropical diseases                            20
         human molecular genetics                                    19
         nature communications                                       19
         bmc public health                                           15
         movement disorders                                          15
         brain                                                       14
         journal of neuroscience                                     13
         biochemical journal                                         12
         developmental cell                                          12
         current biology                                             11
         journal of general virology                                 11
         bmj                                                         10
         malaria journal                                             10
         plos computational biology                                  10
         development                                                  9
         cell reports                                                 9
         journal of virology                                          9
         plosone                                                      9
         neuron                                                       9
         bmj open                                                     9
         molecular microbiology                                       8
         proceedings of the royal society b: biological sciences      8
         journal of physiology                                        8
                                                                   ...
```

```
theranostics                                                 1
journal of acquired immune deficiency syndromes              1
neurogenetics                                                1
american journal of medical genetics part a                  1
international journal of law in context                      1
population, space and place                                  1
j clin microbiol                                             1
nicotine and tobaco research                                 1
international journal for parasitology                       1
tropical animal health & production                          1
frontiers in developmental psychology                        1
molecular genetics & genomic medicine                        1
expert reviews in anti-infective chemotherapy                1
developmental biology                                        1
international journal of health geographics                  1
vascular pharmacology                                        1
cell press - cell reports                                    1
frontiers in immunology                                      1
cerebral cortex print                                        1
genesis: journal of genetics                                 1
j biol chemistry                                             1
clinical transcriptional science                             1
parasit vectors.                                             1
journal of cheminformatics                                   1
maternal and child nutrition                                 1
sci rep                                                      1
future neurology                                             1
tetrahedron letters                                          1
public health ethics                                         1
immnunobiology                                               1
Name: journal_lc, Length: 891, dtype: int64
```

```
In [86]: import re
         cleaning_df[cleaning_df['journal_lc'].str.contains('plos')].head()
```

```
Out[86]:       PMID/PMCID Publisher               Journal title  \
         1278  PMC3744396      PLOS  PLOS Computational Biology
         1279  PMC3681601      PLoS               PLoS Genetics
         1280     3715439      Plos               Plos Genetics
         1281  PMC3493395      PLOS                    PLOS NTD
         1282     3517619      PLoS                    PLoS ONE

                                                Article title  \
         1278  Spike triggered hormone secretion in vasopress...
         1279  Meiosis-specific stable binding of Augmin to a...
         1280  Strabismus promotes recruitment and degradatio...
         1281  Chitinase 3-like 1 protein levels are elevated...
         1282  HCN1 and HCN2 in Rat DRG Neurons: Levels in No...
```

```
     COST (č) charged to Wellcome (inc VAT when charged)  \
1278                                            č1429.13
1279                                            č1494.42
1280                                            č1761.48
1281                                            č1283.76
1282                                            č1001.03

                      journal_lc
1278  plos computational biology
1279               plos genetics
1280               plos genetics
1281                    plos ntd
1282                    plos one
```

1.3 cleaning "biology" journals

```
In [87]: # cleaning_df[cleaning_df['journal_lc'].str.contains('biol ')]['journal_lc'].value_coun
         cleaning_df[cleaning_df['journal_lc'].str.contains('biol')]['journal_lc'].value_counts(
```

```
Out[87]: journal of biological chemistry
         current biology
         plos computational biology
         proceedings of the royal society b: biological sciences
         molecular microbiology
         neurobiology of aging
         journal of structural biology
         the journal of biological chemistry
         cellular microbiology
         biology open
         j biol chem.
         acs chemical biology
         fems microbiology letters
         advances in experimental medicine and biology
         biological psychiatry
         journal of molecular biology
         biology letters
         genome biology
         matrix biology
         microbiology
         rna biology
         jnl biological chemistry
         journal of leukocyte biology
         molecular biology
         acta crystallographica section f: structural biology and crystallization communications
         plos biology
         molecular biology and evolution
         glycobiology
```

```
journal of clinical microbiology
current opinion microbiology

journal of biol chem
molecular membrane biology
biology
curr biol.
international journal of biochemistry & cell biology
the journal of steroid biochemistry & molecular biology
studies in history and philosophy of science part c: studies in history and philosophy
journal of biological physics
methods in molecular biology
diagnostic microbiology and infectious disease
biol open
j steroid biochem mol biol
chemsitry & biology
free radical and biology medicine
developmental biology
neurobiology of disease
biological psychology
the international journal of biochemistry & cell biology
bmc molecular biology
frontiers in t cell biology
philosophical transactions of the royal society of london. series b, biological science
reproductive medicine and biology
biological chemistry
j clin microbiol
bmc biology
molecular and cellular biology
neuropsychobiology
future microbiology
acta crystallographica section d: biological crystallography
bmc genome biology
Name: journal_lc, Length: 84, dtype: int64
```

```
In [88]: cleaning_df['journal_lc'] = cleaning_df['journal_lc'].str.replace('biol ', 'biology')
         cleaning_df['journal_lc'] = cleaning_df['journal_lc'].str.replace('biol.', 'biology')
         cleaning_df['journal_lc'] = cleaning_df['journal_lc'].str.replace('biologygy', 'biology
         cleaning_df['journal_lc'] = cleaning_df['journal_lc'].str.replace('biologygical', 'biol

         cleaning_df[cleaning_df['journal_lc'].str.contains('biol')]['journal_lc'].value_counts(
```

```
Out[88]: journal of biological chemistry
         current biology
         plos computational biology
         proceedings of the royal society b: biological sciences
         journal of structural biology
         molecular microbiology
```

```
neurobiology of aging
the journal of biological chemistry
acs chemical biology
biology open
cellular microbiology
j biologychem.
journal of molecular biology
biology letters
fems microbiology letters
biological psychiatry
advances in experimental medicine and biology
matrix biology
journal of leukocyte biology
rna biology
microbiology
jnl biological chemistry
genome biology
journal of clinical microbiology
acta crystallographica section f: structural biology and crystallization communications
glycobiology
molecular biology and evolution
plos biology
molecular biology
photochemical & photobiological sciences

diagnostic microbiology and infectious disease
biologyopen
mol biologyand evolution
molecular membrane biology
biology
journal of biologychem
international journal of biochemistry & cell biology
journal of evolutionary biology
the journal of steroid biochemistry & molecular biology
acta crystallographica section d, biological crystallography
journal of microbiology
trends in microbiology
journal of biological physics
methods in molecular biology
j steroid biochem mol biol
acta crystallographica section d: biological crystallography
free radical and biology medicine
developmental biology
biological psychology
the international journal of biochemistry & cell biology
bmc molecular biology
philosophical transactions of the royal society of london. series b, biological science
reproductive medicine and biology
```

```
         biological chemistry
         european journal of cell biology
         bmc biology
         molecular and cellular biology
         neuropsychobiology
         future microbiology
         bmc genome biology
         Name: journal_lc, Length: 84, dtype: int64
```

```
In [92]: cleaning_df[cleaning_df['journal_lc'].str.contains('acta')]['journal_lc'].value_counts(
```

```
Out[92]: acta neuropathologica
         acta crystallographica section f: structural biology and crystallization communications
         biochimica et bioohysica acta - molecular basis of disease
         acta diabetologica
         acta crystallographica section d, biological crystallography
         biochimica et biophysica acta - molecular basis of disease
         acta d
         acta crystallography d
         acta crystallographica, section d
         acta neuropathol
         biochimica et biophysica acta (bba) - molecular cell research
         acta crystallographica section d: biological crystallography
         acta dermato venereologica
         acta physiol
         acta f
         acta opthalmologica
         biochimica et bioohysica acta - gene regulatory mechanisms
         Name: journal_lc, dtype: int64
```

1.4 cleaning prices

```
In [95]: cleaning_df['clean_price'] = cleaning_df['COST (č) charged to Wellcome (inc VAT when ch
         cleaning_df.head(1)
```

```
Out[95]:   PMID/PMCID Publisher           Journal title  \
         0        NaN       CUP  Psychological Medicine   

                                                Article title  \
         0  Reduced parahippocampal cortical thickness in ...   

           COST (č) charged to Wellcome (inc VAT when charged)              journal_lc  \
         0                                              č0.00   psychological medicine   

           clean_price
         0       č0.00
```

```
In [102]: cleaning_df['clean_price'] = cleaning_df['clean_price'].str.replace('č', '')
          cleaning_df['clean_price'] = cleaning_df['clean_price'].str.replace('$', '')
          cleaning_df['clean_price'] = cleaning_df['clean_price'].str.replace(' ', '')
          cleaning_df['clean_price'] = cleaning_df['clean_price'].str.strip()
          cleaning_df.head(1)
          cleaning_df['clean_price'] = cleaning_df['clean_price'].astype('float')
```

```
In [100]: cleaning_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 2126 entries, 0 to 2126
Data columns (total 7 columns):
PMID/PMCID                                             1928 non-null object
Publisher                                              2126 non-null object
Journal title                                          2126 non-null object
Article title                                          2126 non-null object
COST (č) charged to Wellcome (inc VAT when charged)    2126 non-null object
journal_lc                                             2126 non-null object
clean_price                                            2126 non-null object
dtypes: object(7)
memory usage: 132.9+ KB
```

```
In [147]: # cleaning_df[cleaning_df['journal_lc'] == 'plos one'].groupby('journal_lc').mean()
          #getting individual journal means

          means = cleaning_df.groupby('journal_lc').mean()
          stds = cleaning_df.groupby('journal_lc').std(ddof=0) #ddof=0?
          medians = cleaning_df.groupby('journal_lc').median()
```

```
In [143]: cleaning_df.head()
```

```
Out[143]:               PMID/PMCID Publisher           Journal title  \
          0                    NaN       CUP  Psychological Medicine
          1             PMC3679557       ACS       Biomacromolecules
          2  23043264  PMC3506128        ACS              J Med Chem
          3    23438330 PMC3646402       ACS              J Med Chem
          4   23438216 PMC3601604        ACS              J Org Chem

                                                 Article title  \
          0  Reduced parahippocampal cortical thickness in ...
          1  Structural characterization of a Model Gram-ne...
          2  Fumaroylamino-4,5-epoxymorphinans and related ...
          3  Orvinols with mixed kappa/mu opioid receptor a...
          4  Regioselective opening of myo-inositol orthoes...

            COST (č) charged to Wellcome (inc VAT when charged)              journal_lc  \
          0                                              č0.00   psychological medicine
          1                                           č2381.04        biomacromolecules
          2                                            č642.56               j med chem
          3                                            č669.64               j med chem
```

```
4                                            č685.88             j org chem

   clean_price
0         0.00
1      2381.04
2       642.56
3       669.64
4       685.88
```

```
In [144]: means.head(3)
```

```
Out[144]:                                    clean_price
          journal_lc
          academy of nutrition and dietetics     2379.540
          acs chemical biology                   1418.186
          acs chemical neuroscience              1186.800
```

```
In [145]: medians.head(3)
```

```
Out[145]:                                    clean_price
          journal_lc
          academy of nutrition and dietetics      2379.54
          acs chemical biology                    1294.59
          acs chemical neuroscience               1186.80
```

```
In [148]: stds.head(3)
```

```
Out[148]:                                    clean_price
          journal_lc
          academy of nutrition and dietetics     0.000000
          acs chemical biology                 453.751465
          acs chemical neuroscience              0.000000
```