Data Wrangling Practice: Armenian online job market postings

This projects reviews and cleans a dataset of 19,000 online job posts from 2004 to 2015 that were posted through an Armenian human resource portal. It is hosted on Kaggle. From the Kaggle page: The online job market is a good indicator of overall demand for labor in an economy. This dataset consists of 19,000 job postings from 2004 to 2015 posted on CareerCenter, an Armenian human resource portal. Since postings are text documents and tend to have similar structures, text mining can be used to extract features like posting date, job title, company name, job description, salary, and more. Postings that had no structure or were not job-related were removed. The data was originally scraped from a Yahoo! mailing group.

Gather

In [1]:

```
import pandas as pd
import zipfile
```

In [2]:

```
# Extract all contents from zip file
with zipfile.ZipFile('armenian-online-job-postings.zip', 'r') as myzip:
    myzip.extractall()
```

In [3]:

```
# Read CSV (comma-separated) file into DataFrame
df = pd.read_csv('online-job-postings.csv')
```

Assess

In [4]:

```
df
```

Out[4]:

	jobpost	date	Title	Company	AnnouncementCode	Term	Elig
0	AMERIA Investment Consulting Company\r\nJOB TI...	Jan 5, 2004	Chief Financial Officer	AMERIA Investment Consulting Company	NaN	NaN	NaN
1	International Research & Exchanges Board (IREX...	Jan 7, 2004	Full-time Community Connections Intern (paid i...	International Research & Exchanges Board (IREX)	NaN	NaN	NaN
2	Caucasus Environmental NGO Network (CENN)\r\nJ...	Jan 7, 2004	Country Coordinator	Caucasus Environmental NGO Network (CENN)	NaN	NaN	NaN
3	Manoff Group\r\nJOB TITLE: BCC Specialist\r\n...	Jan 7, 2004	BCC Specialist	Manoff Group	NaN	NaN	NaN
4	Yerevan Brandy Company\r\nJOB TITLE: Software...	Jan 10, 2004	Software Developer	Yerevan Brandy Company	NaN	NaN	NaN
5	Boutique "Appollo"\r\nJOB TITLE: Saleswoman\r...	Jan 10,	Saleswoman	Boutique "Appollo"	NaN	NaN	NaN

	jobpost	2004 date	Title	Company	AnnouncementCode	Term	Elig
6	OSI Assistance Foundation - Armenian Branch Of...	Jan 11, 2004	Chief Accountant/ Finance Assistant	OSI Assistance Foundation - Armenian Branch Of...	NaN	NaN	NaN
7	International Research & Exchanges Board (IREX...	Jan 13, 2004	Non-paid part or full time Programmatic Intern	International Research & Exchanges Board (IREX)	NaN	NaN	NaN
8	Yerevan Brandy Company \r\nJOB TITLE: Assista...	Jan 13, 2004	Assistant to Managing Director	Yerevan Brandy Company	NaN	NaN	NaN
9	American Embassy Yerevan\r\nANNOUNCEMENT NUMBE...	Jan 13, 2004	Program Assistant (INL), FSN-8; FP-6*	American Embassy Yerevan\r\nANNOUNCEMENT NUMBE...	NaN	NaN	NaN
10	International Research & Exchanges Board (IREX...	Jan 13, 2004	Short-Term Travel Grants (STG) Program	International Research & Exchanges Board (IREX)	NaN	NaN	NaN
11	International Research & Exchanges Board (IREX...	Jan 13, 2004	Non-paid part or full time Administrative Intern	International Research & Exchanges Board (IREX)	NaN	NaN	NaN
12	Institute for Sustainable Communities (ISC)\r\...	Jan 13, 2004	Chief of Party (COP)	Institute for Sustainable Communities (ISC)	NaN	NaN	NaN
13	Food Security Regional Cooperation and Stabili...	Jan 14, 2004	Community Development, Capacity Building and C...	Food Security Regional Cooperation and Stabili...	NaN	NaN	NaN
14	Teleplus LLC\r\nJOB TITLE: General Manager\r\...	Jan 14, 2004	General Manager	Teleplus LLC	NaN	NaN	NaN
15	NetCall Communications\r\nJOB TITLE: Network ...	Jan 15, 2004	Network Administrator	NetCall Communications	NaN	NaN	NaN
16	SOC.Stockholm\r\nTITLE: Utopian World Champi...	Jan 15, 2004	Utopian World Championship 2004	SOC.Stockholm	NaN	NaN	NaN
17	United Nations Development Programme, Armenia\...	Jan 15, 2004	Country Economist (NOB)	United Nations Development Programme, Armenia	NaN	NaN	NaN
18	Counterpart International, Inc.\r\nCounterpart...	Jan 16, 2004	Driver/ Logistics Assistant	Counterpart International, Inc.\r\nCounterpart...	NaN	NaN	NaN
19	Xalt LLC\r\nJOB TITLE: Graphic Designer\r\nPO...	Jan 16, 2004	Graphic Designer	Xalt LLC	NaN	NaN	NaN
20	CUTS Centre for International Trade, Economics...	Jan 19, 2004	Administrative Assistant	CUTS Centre for International Trade, Economics...	NaN	NaN	NaN
21	Yerevan Brandy Company\r\nJOB TITLE: Lawyer\r...	Jan 19, 2004	Lawyer	Yerevan Brandy Company	NaN	NaN	NaN

	jobpost	date	Title	Company	AnnouncementCode	Term	Elig
22	ACDI/VOCA\r\nJOB TITLE: Marketing Advisor\r\n...	Jan 19, 2004	Marketing Advisor	ACDI/VOCA	NaN	NaN	NaN
23	Armenia TV\r\nJOB TITLE: Chief/ Supervisor of...	Jan 20, 2004	Chief/ Supervisor of Programs Department	Armenia TV	NaN	NaN	NaN
24	Media Diversity Institute\r\nJOB TITLE: Journ...	Jan 20, 2004	Journalism Trainer	Media Diversity Institute	NaN	NaN	NaN
25	Armenia School Connectivity Program\r\nJOB TIT...	Jan 21, 2004	Deputy Program Director	Armenia School Connectivity Program	NaN	NaN	NaN
26	Yerevan State University (YSU) and Civic Educa...	Jan 21, 2004	Student Forum: Student Conference and Debate F...	Yerevan State University (YSU) and Civic Educa...	NaN	NaN	NaN
27	Media Diversity Institute\r\nJOB TITLE: Repor...	Jan 21, 2004	Reporting Diversity Workshop for Journalists	Media Diversity Institute	NaN	NaN	NaN
28	Media Diversity Institute\r\nTITLE: Reportin...	Jan 21, 2004	Reporting Diversity Workshop for Journalists	Media Diversity Institute	NaN	NaN	NaN
29	SEF International\r\nJOB TITLE: Chief Account...	Jan 23, 2004	Chief Accountant	SEF International	NaN	NaN	NaN
...	...	...	...	...	...	...	...
18971	Career Center NGO\r\n\r\n\r\nTITLE: Reception...	Dec 22, 2015	Receptionist/ Administrative Assistant	Career Center NGO	NaN	Part-time/ Full-time	Everyon
18972	Essential Solutions LLC\r\n\r\n\r\nTITLE: Jav...	Dec 22, 2015	Java, Grails/ Groovie Developer	Essential Solutions LLC	NaN	NaN	NaN
18973	Essential Solutions LLC\r\n\r\n\r\nTITLE: Sen...	Dec 22, 2015	Senior Java Software Engineer	Essential Solutions LLC	NaN	NaN	NaN
18974	LTX-Credence Armenia LLC\r\n\r\n\r\nTITLE: CA...	Dec 23, 2015	CAD Librarian (Printed Circuit Board Design)	LTX-Credence Armenia LLC	NaN	NaN	NaN
18975	LTX-Credence Armenia LLC\r\n\r\n\r\nTITLE: In...	Dec 23, 2015	Intern (Coop Tech): Precision Analog HW Design	LTX-Credence Armenia LLC	NaN	NaN	NaN
18976	SAS Group LLC\r\n\r\n\r\nTITLE: Senior Financ...	Dec 22, 2015	Senior Financial Specialist	SAS Group LLC	NaN	NaN	NaN
18977	Strategic Development Agency (SDA) NGO\r\n\r\n...	Dec 22, 2015	Business Consultant	Strategic Development Agency (SDA) NGO	NaN	NaN	NaN

	jobpost	date	Title	Company	AnnouncementCode	Term	Elig
18978	Representation of the "French Office for Immig...	Dec 22, 2015	National Expert on Socio-Demographic Statistics	Representation of the "French Office for Immig...	NaN	NaN	NaN
18979	Care Building Services LLC\r\n\r\n\r\nTITLE: ...	Dec 23, 2015	Project Assistant	Care Building Services LLC	NaN	NaN	All eligib candidat
18980	LTX-Credence Armenia LLC\r\n\r\n\r\nTITLE: Ha...	Dec 23, 2015	Hardware Engineer	LTX-Credence Armenia LLC	NaN	NaN	NaN
18981	BDO Armenia CJSC\r\n\r\n\r\nTITLE: Auditor\r\...	Dec 23, 2015	Auditor	BDO Armenia CJSC	NaN	NaN	NaN
18982	BDO Armenia CJSC\r\n\r\n\r\nTITLE: Senior Aud...	Dec 23, 2015	Senior Auditor	BDO Armenia CJSC	NaN	NaN	NaN
18983	Care Building Services LLC\r\n\r\n\r\nTITLE: ...	Dec 23, 2015	Construction Site Manager	Care Building Services LLC	NaN	NaN	All eligib candidat
18984	Care Building Services LLC\r\n\r\n\r\nTITLE: ...	Dec 23, 2015	Electrical Engineer	Care Building Services LLC	NaN	NaN	All eligib candidat
18985	Strategic Development Agency (SDA) NGO\r\n\r\n...	Dec 22, 2015	Project Assistant	Strategic Development Agency (SDA) NGO	NaN	NaN	NaN
18986	Strategic Development Agency (SDA) NGO\r\n\r\n...	Dec 22, 2015	Public Relations Specialist	Strategic Development Agency (SDA) NGO	NaN	NaN	NaN
18987	PicsArt LLC\r\n\r\n\r\nTITLE: Junior Mobile V...	Dec 24, 2015	Junior Mobile Visual UI/ UX Designer	PicsArt LLC	NaN	NaN	NaN
18988	Berg Development LLC\r\n\r\n\r\nTITLE: JavaSc...	Dec 24, 2015	JavaScript Professional	Berg Development LLC	NaN	Full-time	NaN
18989	"Transport PIU" State Institution of the RA Mi...	Dec 24, 2015	Leading Specialist/ Accountant of Financial an...	"Transport PIU" State Institution of the RA Mi...	NaN	NaN	All intere and qua candidat
18990	NASDAQ OMX Armenia OJSC\r\n\r\n\r\nTITLE: Jav...	Dec 24, 2015	Java Developer	NASDAQ OMX Armenia OJSC	NaN	Full-time	NaN
18991	NASDAQ OMX Armenia OJSC\r\n\r\n\r\nTITLE: C/ ...	Dec 24, 2015	C/ C++ Developer	NASDAQ OMX Armenia OJSC	NaN	Full-time	NaN
18992	Macadamian AR CJSC\r\n\r\n\r\nTITLE: .NET Dev...	Dec 25, 2015	.NET Developer	Macadamian AR CJSC	NaN	NaN	NaN
18993	"Coca-Cola Hellenic Bottling Company Armenia" ...	Dec 25, 2015	Recruitment Specialist	"Coca-Cola Hellenic Bottling Company Armenia" ...	NaN	Full-time	All intere candidat
18994	"Coca-Cola Hellenic Bottling Company Armenia" ...	Dec 25, 2015	Capability Development	"Coca-Cola Hellenic Bottling Company Armenia" ...	NaN	NaN	All intere candidat

	jobpost	date	Title	Company	AnnouncementCode	Term	Elig
18995	"Transport PIU" State Institution of the RA Mi...	Dec 24, 2015	Deputy Director	"Transport PIU" State Institution of the RA Mi...	NaN	NaN	All intere and qua candidat
18996	Technolinguistics NGO\r\n\r\n\r\nTITLE: Senio...	Dec 28, 2015	Senior Creative UX/ UI Designer	Technolinguistics NGO	NaN	Full-time	NaN
18997	"Coca-Cola Hellenic Bottling Company Armenia" ...	Dec 30, 2015	Category Development Manager	"Coca-Cola Hellenic Bottling Company Armenia" ...	NaN	Full-time	All intere professi
18998	"Coca-Cola Hellenic Bottling Company Armenia" ...	Dec 30, 2015	Operational Marketing Manager	"Coca-Cola Hellenic Bottling Company Armenia" ...	NaN	Full-time	All intere professi
18999	San Lazzaro LLC\r\n\r\n\r\nTITLE: Head of O...	Dec 30, 2015	Head of Online Sales Department	San Lazzaro LLC	NaN	NaN	NaN
19000	"Kamurj" UCO CJSC\r\n\r\n\r\nTITLE: Lawyer in...	Dec 30, 2015	Lawyer in Legal Department	"Kamurj" UCO CJSC	NaN	Full-time	NaN

19001 rows × 24 columns

In [5]:

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 19001 entries, 0 to 19000
Data columns (total 24 columns):
jobpost             19001 non-null object
date                19001 non-null object
Title               18973 non-null object
Company             18994 non-null object
AnnouncementCode    1208 non-null object
Term                7676 non-null object
Eligibility         4930 non-null object
Audience            640 non-null object
StartDate           9675 non-null object
Duration            10798 non-null object
Location            18969 non-null object
JobDescription      15110 non-null object
JobRequirment       16479 non-null object
RequiredQual        18517 non-null object
Salary              9623 non-null object
ApplicationP        18941 non-null object
OpeningDate         18295 non-null object
Deadline            18936 non-null object
Notes               2211 non-null object
AboutC              12470 non-null object
Attach              1559 non-null object
Year                19001 non-null int64
Month               19001 non-null int64
IT                  19001 non-null bool
dtypes: bool(1), int64(2), object(21)
memory usage: 3.4+ MB
```

- Missing values (NaN)
- StartDate inconsistencies (ASAP)
- Fix nondescriptive column headers (ApplicationP, AboutC, RequiredQual ... and also JobRequirment)

Clean

Define

- Select all nondescriptive and misspelled column headers (ApplicationP, AboutC, RequiredQual, JobRequirment) and replace them with full words (ApplicationProcedure, AboutCompany, RequiredQualifications, JobRequirement)

- Select all records in the StartDate column that have "As soon as possible", "Immediately", etc. and replace the text in those cells with "ASAP"

Code

In [6]:

```
df_clean = df.copy()
```

- Select all nondescriptive and misspelled column headers (ApplicationP, AboutC, RequiredQual, JobRequirment) and replace them with full words (ApplicationProcedure, AboutCompany, RequiredQualifications, JobRequirement)

In [7]:

```
df_clean = df_clean.rename(columns={'ApplicationP': 'ApplicationProcedure',
                                    'AboutC': 'AboutCompany',
                                    'RequiredQual': 'RequiredQualifications',
                                    'JobRequirment': 'JobRequirements'})
```

- Select all records in the StartDate column that have "As soon as possible", "Immediately", etc. and replace the text in those cells with "ASAP"

In [8]:

```
asap_list = ['Immediately', 'As soon as possible', 'Upon hiring',
             'Immediate', 'Immediate employment', 'As soon as possible.', 'Immediate job opportunit
y',
             '"Immediate employment, after passing the interview."',
             'ASAP preferred', 'Employment contract signature date',
             'Immediate employment opportunity', 'Immidiately', 'ASA',
             'Asap', '"The position is open immediately but has a flexible start date depending on
the candidates earliest availability."',
             'Immediately upon agreement', '20 November 2014 or ASAP',
             'immediately', 'Immediatelly',
             '"Immediately upon selection or no later than November 15, 2009."',
             'Immediate job opening', 'Immediate hiring', 'Upon selection',
             'As soon as practical', 'Immadiate', 'As soon as posible',
             'Immediately with 2 months probation period',
             '12 November 2012 or ASAP', 'Immediate employment after passing the interview',
             'Immediately/ upon agreement', '01 September 2014 or ASAP',
             'Immediately or as per agreement', 'as soon as possible',
             'As soon as Possible', 'in the nearest future', 'immediate',
             '01 April 2014 or ASAP', 'Immidiatly', 'Urgent',
             'Immediate or earliest possible', 'Immediate hire',
             'Earliest  possible', 'ASAP with 3 months probation period.',
             'Immediate employment opportunity.', 'Immediate employment.',
             'Immidietly', 'Imminent', 'September 2014 or ASAP', 'Imediately']

for phrase in asap_list:
    df_clean.StartDate.replace(phrase, 'ASAP', inplace=True)
```

Test

In [9]:

```
df_clean.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 19001 entries, 0 to 19000
Data columns (total 24 columns):
jobpost                  19001 non-null object
date                     19001 non-null object
Title                    18973 non-null object
Company                  18994 non-null object
AnnouncementCode         1208 non-null object
Term                     7676 non-null object
Eligibility              4930 non-null object
Audience                 640 non-null object
StartDate                9675 non-null object
Duration                 10798 non-null object
```

```
Duration                  10796 non-null object
Location                  18969 non-null object
JobDescription            15110 non-null object
JobRequirements           16479 non-null object
RequiredQualifications    18517 non-null object
Salary                    9623 non-null object
ApplicationProcedure      18941 non-null object
OpeningDate               18295 non-null object
Deadline                  18936 non-null object
Notes                     2211 non-null object
AboutCompany              12470 non-null object
Attach                    1559 non-null object
Year                      19001 non-null int64
Month                     19001 non-null int64
IT                        19001 non-null bool
dtypes: bool(1), int64(2), object(21)
memory usage: 3.4+ MB
```

In [10]:

```
df_clean.StartDate.value_counts()
```

Out[10]:

```
ASAP                                                6856
01 September 2012                                     31
March 2006                                            27
November 2006                                         22
January 2010                                          19
February 2014                                         17
01 February 2005                                      17
February 2011                                         16
September 2010                                        16
TBD                                                   16
March 2011                                            15
September 2008                                        15
February 2007                                         14
01 February 2015                                      14
01 July 2014                                          14
01 March 2006                                         13
01 September 2010                                     13
01 September 2008                                     13
January 2016                                          13
01 April 2014                                         12
01 October 2008                                       12
September 2013                                        12
01 October 2012                                       12
01 March 2012                                         12
June 2008                                             12
Fall 2010                                             12
01 April 2012                                         11
01 September 2004                                     11
01 February 2012                                      11
January 2011                                          11
                                                    ...
10 May 2007                                            1
20 November 2009                                       1
Beginning of March 2006                                1
07 May 2012                                            1
20 May 2015                                            1
The start period of trainings is mid November.         1
08 April 2015                                          1
28 July 2008                                           1
08 October 2012                                        1
30 October                                             1
15 November 2004                                       1
10 June 2006                                           1
15 October 2014                                        1
09 September 2013                                      1
21 January 2010 to 25 June 2010                        1
01 March  2009                                         1
17 December 2012                                       1
26 October 2011                                        1
09 August 04                                           1
September  2005                                        1
28 March 2005                                          1
15 February 2012                                       1
```

```
15 February 2012                                    1
08 June 2009                                        1
03 November 2008                                    1
09 April 2012                                       1
23 May 2009, 18:00                                  1
1 October 2007                                      1
15 August 2011                                      1
7 December 2007                                     1
Mid-December 2007                                   1
Name: StartDate, Length: 1140, dtype: int64
```

In [ ]:

```
for phrase in asap_list:
    assert phrase not in df_clean.StartDate.values
```

In [ ]:

```
import pdfkit
pdfkit.from_file('online-job-postings.html', 'online-job-postings.pdf')
```