StackOverFlow Survey

October 27, 2021

[15]:
```
import gzip
import math
import random
from collections import defaultdict
import copy
import datetime
from matplotlib import pyplot as plt
import numpy as np
import random
from tqdm.notebook import tqdm
import zipfile
import pandas as pd
```

[17]:
```
# find files
import os
def list_file_path(root = './',ext = ''):
    for root, dirs, files in os.walk(root):
        for name in files:
            if ext in name:
                print(os.path.abspath(os.path.join(root, name)))
list_file_path(ext='zip')
```

```
/home/yel004/CSE258/archive.zip
/home/yel004/CSE258/project_data/stack-overflow-developer-survey-2021.zip
```

[29]:
```
list_file_path(root = './project_data')
```

```
/home/yel004/CSE258/project_data/so_survey_2021.pdf
/home/yel004/CSE258/project_data/survey_results_schema.csv
/home/yel004/CSE258/project_data/survey_results_public.csv
/home/yel004/CSE258/project_data/stack-overflow-developer-survey-2021.zip
/home/yel004/CSE258/project_data/survey_results_responses.csv
/home/yel004/CSE258/project_data/survey_results_questions.csv
/home/yel004/CSE258/project_data/README_2021.txt
```

[28]:
```
#kaggle dataset
path_to_zip_file = "archive.zip"
directory_to_extract_to = "/home/yel004/CSE258/project_data"
```

```
with zipfile.ZipFile(path_to_zip_file, 'r') as zip_ref:
    zip_ref.extractall(directory_to_extract_to)
```

1 select dataset and decision on choosing files

The Kaggle dataset is missing the readme file. By searching the source in Kaggle, I found the original dataset which includs a readme and formatted pdf of the survey. (Link: https://insights.stackoverflow.com/survey). However,Kaggle data has some preprocessing done beforehand. The main different is still with in the question files.

1.1 StackOverFlow Original Schema files

[19]:
```
scheme_path = "/home/yel004/CSE258/project_data/survey_results_schema.csv"
schema = pd.read_csv(scheme_path)
```

[24]:
```
schema.head()
```

[24]:
```
     qid       qname                                              question  \
0  QID16          S0  <div><span style="font-size:19px;"><strong>Hel…
1  QID12    MetaInfo                                  Browser Meta Info
2   QID1          S1  <span style="font-size:22px; font-family: aria…
3   QID2  MainBranch  Which of the following options best describes …
4  QID24  Employment  Which of the following best describes your cur…

   force_resp  type selector
0       False    DB       TB
1       False  Meta  Browser
2       False    DB       TB
3        True    MC     SAVR
4       False    MC     MAVR
```

[27]:
```
schema['question'][0]
```

[27]:
```
'<div><span style="font-size:19px;"><strong>Hello world!
</strong></span></div>\n\n<div> </div>\n\n<div>Thank you for taking the 2021
Stack Overflow Developer Survey, the longest running survey of software
developers (and anyone else who codes!) on Earth. </div>\n\n<div>
</div>\n\n<div>As in previous years, anonymized results of the survey will be
made publicly available under the Open Database License, where anyone can
download and analyze the data. On that note, throughout the survey, certain
answers you and your peers give will be treated as personally identifiable
information, and therefore kept out of the anonymized results file. We\'ll call
out each of those in the survey with a note saying "This information will be
kept private." </div>\n\n<div> </div>\n\n<div>There are six sections in this
survey. The 2nd, 3rd, and 4th sections will appear in a random
order.</div><div><br></div>\n\n<div>   1. Basic Information</div>\n\n<div>   2.
```

```
Education, Work, and Career</div>\n\n<div>   3. Technology and Tech
Culture</div>\n\n<div>   4. Stack Overflow Usage + Community</div>\n\n<div>   5.
Demographic Information </div>\n\n<div>   6. Final Questions</div>\n\n<div>
\n<div>Most questions in this survey are optional. Required questions are marked
with *. This anonymous survey will take about 10 minutes to complete. We
encourage you to complete it in one
sitting.</div><div><br></div>\n</div>\n\n<div><strong>If you use security or ad-
blocking plugins, you may see error messages</strong></div>\n\n<div>Our third-
party software provider, Qualtrics, does not work well with certain ad blockers
and security software. To avoid error messages that prevent you from taking the
survey, please try specifically unblocking Qualtrics in your plugin or pausing
the plugin while you take the survey. </div>\n\n<div> </div>\n\n<div>To begin,
click <strong>Next.</strong></div>'
```

1.2 StackOverFlow Original Response file

[33]:
```
stack_Response_path = "/home/yel004/CSE258/project_data/survey_results_public.
↪csv"
stack_Response = pd.read_csv(stack_Response_path)
```

[34]:
```
stack_Response.head()
```

[34]:
```
   ResponseId                                     MainBranch  \
0           1                  I am a developer by profession
1           2          I am a student who is learning to code
2           3  I am not primarily a developer, but I write co…
3           4                  I am a developer by profession
4           5                  I am a developer by profession

                                      Employment  \
0  Independent contractor, freelancer, or self-em…
1                             Student, full-time
2                             Student, full-time
3                             Employed full-time
4  Independent contractor, freelancer, or self-em…

                                        Country US_State UK_Country  \
0                                      Slovakia      NaN        NaN
1                                   Netherlands      NaN        NaN
2                            Russian Federation      NaN        NaN
3                                       Austria      NaN        NaN
4  United Kingdom of Great Britain and Northern I…      NaN    England

                                        EdLevel     Age1stCode  \
0  Secondary school (e.g. American high school, G…  18 - 24 years
1       Bachelor's degree (B.A., B.S., B.Eng., etc.)  11 - 17 years
```

```
2       Bachelor's degree (B.A., B.S., B.Eng., etc.)  11 - 17 years
3    Master's degree (M.A., M.S., M.Eng., MBA, etc.)  11 - 17 years
4    Master's degree (M.A., M.S., M.Eng., MBA, etc.)   5 - 10 years

                                            LearnCode YearsCode  …  \
0  Coding Bootcamp;Other online resources (ex: vi…       NaN  …
1  Other online resources (ex: videos, blogs, etc…         7  …
2  Other online resources (ex: videos, blogs, etc…       NaN  …
3                                               NaN       NaN  …
4                          Friend or family member        17  …

               Age Gender Trans                Sexuality  \
0  25-34 years old    Man    No  Straight / Heterosexual
1  18-24 years old    Man    No  Straight / Heterosexual
2  18-24 years old    Man    No        Prefer not to say
3  35-44 years old    Man    No  Straight / Heterosexual
4  25-34 years old    Man    No                      NaN

                      Ethnicity               Accessibility  \
0  White or of European descent            None of the above
1  White or of European descent            None of the above
2            Prefer not to say            None of the above
3  White or of European descent  I am deaf / hard of hearing
4  White or of European descent            None of the above

        MentalHealth           SurveyLength                   SurveyEase  \
0  None of the above  Appropriate in length                         Easy
1  None of the above  Appropriate in length                         Easy
2  None of the above  Appropriate in length                         Easy
3               NaN  Appropriate in length  Neither easy nor difficult
4               NaN  Appropriate in length                         Easy

  ConvertedCompYearly
0             62268.0
1                 NaN
2                 NaN
3                 NaN
4                 NaN

[5 rows x 48 columns]
```

[35]:
```
stack_Response.iloc[0]
```

[35]:
```
ResponseId                                                      1
MainBranch                                  I am a developer by
profession
```

```
Employment                      Independent contractor, freelancer, or self-
em…
Country
Slovakia
US_State
NaN
UK_Country
NaN
EdLevel                         Secondary school (e.g. American high school,
G…
Age1stCode                                                        18 - 24
years
LearnCode                       Coding Bootcamp;Other online resources (ex:
vi…
YearsCode
NaN
YearsCodePro
NaN
DevType                                                          Developer,
mobile
OrgSize                                                         20 to 99
employees
Currency                                                         EUR European
Euro
CompTotal
4800.0
CompFreq
Monthly
LanguageHaveWorkedWith
C++;HTML/CSS;JavaScript;Objective-C;PHP;Swift
LanguageWantToWorkWith
Swift
DatabaseHaveWorkedWith
PostgreSQL;SQLite
DatabaseWantToWorkWith
SQLite
PlatformHaveWorkedWith
NaN
PlatformWantToWorkWith
NaN
WebframeHaveWorkedWith
Laravel;Symfony
WebframeWantToWorkWith
NaN
MiscTechHaveWorkedWith
NaN
MiscTechWantToWorkWith
```

```
NaN
ToolsTechHaveWorkedWith
NaN
ToolsTechWantToWorkWith
NaN
NEWCollabToolsHaveWorkedWith
PHPStorm;Xcode
NEWCollabToolsWantToWorkWith
Atom;Xcode
OpSys
MacOS
NEWStuck                        Call a coworker or friend;Visit Stack
Overflow…
NEWSOSites                                                   Stack
Overflow
SOVisitFreq                                          Multiple times per
day
SOAccount
Yes
SOPartFreq                                    A few times per month or
weekly
SOComm                                                       Yes,
definitely
NEWOtherComms
No
Age                                                          25-34 years
old
Gender
Man
Trans
No
Sexuality                                             Straight /
Heterosexual
Ethnicity                                          White or of European
descent
Accessibility                                            None of the
above
MentalHealth                                             None of the
above
SurveyLength                                          Appropriate in
length
SurveyEase
Easy
ConvertedCompYearly
62268.0
Name: 0, dtype: object
```

1.3 Kaggle preprocessed question files

[30]:
```
questions_path ="/home/yel004/CSE258/project_data/survey_results_questions.csv"
questions = pd.read_csv(questions_path)
```

[31]:
```
questions.head()
```

[31]:
```
     qid       qname                                                question  \
0   QID2  MainBranch  Which of the following options best describes …
1  QID24  Employment  Which of the following best describes your cur…
2   QID6     Country                                  Where do you live?
3   QID7    US_State  In which state or territory of the USA do you …
4   QID9  UK_Country  In which part of the United Kingdom do you liv…

   force_resp type selector
0        True   MC     SAVR
1       False   MC     MAVR
2        True   MC       DL
3       False   MC       DL
4       False   MC       DL
```

[32]:
```
questions['question'][0]
```

[32]:
```
'Which of the following options best describes you today? Here, by "developer"
we mean "someone who writes code." '
```

Kaggle preprocessed responses files

[36]:
```
kag_response_path = '/home/yel004/CSE258/project_data/survey_results_responses.
 ↪csv'
kag_response = pd.read_csv(kag_response_path)
```

[37]:
```
kag_response.head()
```

[37]:
```
   ResponseId                                         MainBranch  \
0           1                    I am a developer by profession
1           2            I am a student who is learning to code
2           3  I am not primarily a developer, but I write co…
3           4                    I am a developer by profession
4           5                    I am a developer by profession

                                      Employment  \
0  Independent contractor, freelancer, or self-em…
1                             Student, full-time
2                             Student, full-time
3                             Employed full-time
4  Independent contractor, freelancer, or self-em…
```

```
                                                Country US_State UK_Country  \
0                                              Slovakia      NaN        NaN
1                                           Netherlands      NaN        NaN
2                                    Russian Federation      NaN        NaN
3                                               Austria      NaN        NaN
4  United Kingdom of Great Britain and Northern I…      NaN    England

                                                EdLevel     Age1stCode  \
0  Secondary school (e.g. American high school, G…  18 - 24 years
1       Bachelor's degree (B.A., B.S., B.Eng., etc.)  11 - 17 years
2       Bachelor's degree (B.A., B.S., B.Eng., etc.)  11 - 17 years
3    Master's degree (M.A., M.S., M.Eng., MBA, etc.)  11 - 17 years
4    Master's degree (M.A., M.S., M.Eng., MBA, etc.)   5 - 10 years

                                          LearnCode YearsCode  …  \
0  Coding Bootcamp;Other online resources (ex: vi…       NaN  …
1  Other online resources (ex: videos, blogs, etc…         7  …
2  Other online resources (ex: videos, blogs, etc…       NaN  …
3                                              NaN       NaN  …
4                          Friend or family member        17  …

              Age Gender Trans               Sexuality  \
0  25-34 years old    Man    No  Straight / Heterosexual
1  18-24 years old    Man    No  Straight / Heterosexual
2  18-24 years old    Man    No        Prefer not to say
3  35-44 years old    Man    No  Straight / Heterosexual
4  25-34 years old    Man    No                     NaN

                     Ethnicity               Accessibility  \
0  White or of European descent            None of the above
1  White or of European descent            None of the above
2            Prefer not to say            None of the above
3  White or of European descent  I am deaf / hard of hearing
4  White or of European descent            None of the above

        MentalHealth           SurveyLength                  SurveyEase  \
0  None of the above  Appropriate in length                        Easy
1  None of the above  Appropriate in length                        Easy
2  None of the above  Appropriate in length                        Easy
3               NaN  Appropriate in length  Neither easy nor difficult
4               NaN  Appropriate in length                        Easy

  ConvertedCompYearly
0             62268.0
1                 NaN
2                 NaN
3                 NaN
```

```
4                 NaN

[5 rows x 48 columns]
```

[38]:
```
kag_response.iloc[0]
```

[38]:
```
ResponseId
1
MainBranch                                          I am a developer by
profession
Employment                        Independent contractor, freelancer, or self-
em…
Country
Slovakia
US_State
NaN
UK_Country
NaN
EdLevel                           Secondary school (e.g. American high school,
G…
Age1stCode                                                        18 - 24
years
LearnCode                         Coding Bootcamp;Other online resources (ex:
vi…
YearsCode
NaN
YearsCodePro
NaN
DevType                                                      Developer,
mobile
OrgSize                                                     20 to 99
employees
Currency                                                     EUR European
Euro
CompTotal
4800.0
CompFreq
Monthly
LanguageHaveWorkedWith
C++;HTML/CSS;JavaScript;Objective-C;PHP;Swift
LanguageWantToWorkWith
Swift
DatabaseHaveWorkedWith
PostgreSQL;SQLite
DatabaseWantToWorkWith
SQLite
PlatformHaveWorkedWith
```

NaN
PlatformWantToWorkWith
NaN
WebframeHaveWorkedWith
Laravel;Symfony
WebframeWantToWorkWith
NaN
MiscTechHaveWorkedWith
NaN
MiscTechWantToWorkWith
NaN
ToolsTechHaveWorkedWith
NaN
ToolsTechWantToWorkWith
NaN
NEWCollabToolsHaveWorkedWith
PHPStorm;Xcode
NEWCollabToolsWantToWorkWith
Atom;Xcode
OperatingSystem
MacOS
NEWStuck Call a coworker or friend;Visit Stack
Overflow…
NEWSOSites Stack
Overflow
SOVisitFreq Multiple times per
day
SOAccount
Yes
SOPartFreq A few times per month or
weekly
SOComm Yes,
definitely
NEWOtherComms
No
Age 25-34 years
old
Gender
Man
Trans
No
Sexuality Straight /
Heterosexual
Ethnicity White or of European
descent
Accessibility None of the
above

```
MentalHealth                                                    None of the
above
SurveyLength                                                 Appropriate in
length
SurveyEase
Easy
ConvertedCompYearly
62268.0
Name: 0, dtype: object
```

1.3.1 Zero users without NA. A smart way of handling missing values must be decided

[39]:
```
len(kag_response)
```

[39]:
```
83439
```

[40]:
```
kag_response_naRem = kag_response.dropna()
```

[41]:
```
len(kag_response_naRem)
```

[41]:
```
0
```

[45]:
```
kag_response.columns
```

[45]:
```
Index(['ResponseId', 'MainBranch', 'Employment', 'Country', 'US_State',
       'UK_Country', 'EdLevel', 'Age1stCode', 'LearnCode', 'YearsCode',
       'YearsCodePro', 'DevType', 'OrgSize', 'Currency', 'CompTotal',
       'CompFreq', 'LanguageHaveWorkedWith', 'LanguageWantToWorkWith',
       'DatabaseHaveWorkedWith', 'DatabaseWantToWorkWith',
       'PlatformHaveWorkedWith', 'PlatformWantToWorkWith',
       'WebframeHaveWorkedWith', 'WebframeWantToWorkWith',
       'MiscTechHaveWorkedWith', 'MiscTechWantToWorkWith',
       'ToolsTechHaveWorkedWith', 'ToolsTechWantToWorkWith',
       'NEWCollabToolsHaveWorkedWith', 'NEWCollabToolsWantToWorkWith',
       'OperatingSystem', 'NEWStuck', 'NEWSOSites', 'SOVisitFreq', 'SOAccount',
       'SOPartFreq', 'SOComm', 'NEWOtherComms', 'Age', 'Gender', 'Trans',
       'Sexuality', 'Ethnicity', 'Accessibility', 'MentalHealth',
       'SurveyLength', 'SurveyEase', 'ConvertedCompYearly'],
      dtype='object')
```

2 TODO:

1. Formulate recommending system question
2. Many features are categorical; how to use them?

[ ]:
```

```