

download_WRDS_CIQ_transcripts_others

May 27, 2025

1 Download Other Files from Capital IQ Transcripts

```
[31]: %matplotlib inline
      # Do below if you want interactive matplotlib plot (). You can zoom in / zoom
      ↪out.
      # %matplotlib notebook

      # reloads modules automatically before entering the execution of code typed at
      ↪the IPython prompt
      %load_ext autoreload
      %autoreload 2
```

The autoreload extension is already loaded. To reload it, use:

```
%reload_ext autoreload
```

```
[32]: import wrds, os, re
      import pandas as pd
      from functions_gen import *
```

```
[33]: # setting up options
      pd.set_option('display.memory_usage', 'deep')
      pd.set_option('display.precision', 2)
      pd.set_option('display.width', 240)
      pd.set_option('display.max_rows', 4000)
      pd.options.display.max_columns = None
      pd.options.display.float_format = '{:,.4f}'.format
```

```
[34]: db = wrds.Connection(wrds_username = '#####')
```

Loading library list...

Done

```
[35]: db.list_tables(library='ciq_transcripts')
```

```
[35]: ['ciqtranscript',
      'ciqtranscriptcollectiontype',
      'ciqtranscriptcomponent',
      'ciqtranscriptcomponenttype',
```

```
'ciqtranscriptdelayreason',
'ciqtranscriptdelayreasantype',
'ciqtranscriptperson',
'ciqtranscriptpresentationtype',
'ciqtranscriptspeakertype',
'wrds_transcript_detail',
'wrds_transcript_person']
```

```
[36]: if os.path.exists("../output") ==False:
      os.makedirs("../output")
```

```
[37]: wrds_transcript_detail = db.get_table(library='ciq_transcripts',
      ↪table='wrds_transcript_detail')
```

```
[38]: wrds_transcript_detail.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 1575627 entries, 0 to 75626
Data columns (total 16 columns):
#   Column                                Non-Null Count  Dtype
---  -
0   companyid                             1575627 non-null object
1   keydevid                               1575627 non-null object
2   transcriptid                           1575627 non-null object
3   headline                               1575627 non-null object
4   mostimportantdateutc                   1575627 non-null object
5   mostimportanttimeutc                   1575627 non-null object
6   keydeveventtypeid                     1575627 non-null object
7   keydeveventtypename                   1575627 non-null object
8   companyname                           1574753 non-null object
9   transcriptcollectiontypeid             1575627 non-null int64
10  transcriptcollectiontypename            1575627 non-null object
11  transcriptpresentationtypeid            1575627 non-null int64
12  transcriptpresentationtypename          1575627 non-null object
13  transcriptcreationdate_utc              1575627 non-null object
14  transcriptcreationtime_utc              1575627 non-null object
15  audiolengthsec                         1532987 non-null object
dtypes: int64(2), object(14)
memory usage: 1.7 GB
```

```
[39]: # wrds_transcript_detail.to_pickle("../output/wrds_transcript_detail.pkl",
      ↪compression='zip')
wrds_transcript_detail = pd.read_pickle("../output/wrds_transcript_detail.pkl",
      ↪compression='zip')
```

```
[40]: wrds_transcript_detail.duplicated(subset=['transcriptid']).value_counts()
```

```
[40]: False    1543357
      True      32270
      Name: count, dtype: int64
```

```
[41]: transcripts = wrds_transcript_detail.drop_duplicates(subset=['transcriptid']).
      ↪sort_values(by='transcriptid')
```

```
[42]: transcripts['transcriptid'] = transcripts['transcriptid'].astype(int)
```

```
[43]: transcripts['quantile_10'], bins = pd.qcut(transcripts.transcriptid, q=10,
      ↪precision=0, retbins=True)
      transcripts['quantile_10'].value_counts(sort=False)
```

```
[43]: quantile_10
(107.0, 273691.0]      154336
(273691.0, 556695.0]   154336
(556695.0, 866928.0]   154335
(866928.0, 1196449.0]   154336
(1196449.0, 1509145.0]  154336
(1509145.0, 1783798.0]  154335
(1783798.0, 2038333.0]  154336
(2038333.0, 2391959.0]  154335
(2391959.0, 2736944.0]  154336
(2736944.0, 3175090.0]  154336
Name: count, dtype: int64
```

```
[44]: bins.tolist()
```

```
[44]: [108.0,
      273690.6,
      556694.6000000001,
      866928.4000000001,
      1196448.6,
      1509145.0,
      1783797.6,
      2038333.4000000004,
      2391958.8000000003,
      2736943.6000000006,
      3175090.0]
```

```
[45]: def get_person_sql(start, end, output, filenamesuf):
      df = db.raw_sql(f"select distinct * from ciq_transcripts.
      ↪wrds_transcript_person where transcriptid>{start} and transcriptid<={end}")
      print(f"\n### Downloading transcript person details, start transcriptid is_
      ↪{start}, end transcriptid is {end} ###\n")
      print("# descriptive statistics")
      print(df.info(), "\n")
```

```

    print(df.duplicated(subset=['transcriptid','transcriptcomponentid']).
↪value_counts(),"\n")
    print(df.describe(),"\n")
    print("# compare sample with wrds_transcript_detail")
    print(df.merge(right=wrds_transcript_detail[(wrds_transcript_detail.
↪transcriptid>start)&(wrds_transcript_detail.transcriptid<=end)],
                on=['transcriptid'], indicator=True, how='outer')._merge.
↪value_counts(), "\n")
    print(f"# saving zip compressed pickle as_
↪wrds_transcript_person_{filenamesuf}")
    df.to_pickle(f"{output}/wrds_transcript_person_{filenamesuf}.pkl",_
↪compression="zip")

```

```

[46]: bins_test = [100, 500, 1000]
i=0
while i <len(bins.tolist())-1:
    print(f"\n##### PROCESSING BATCH {str(i+1).zfill(2)} #####\n")
    start = bins.tolist()[i]
    end = bins.tolist()[i+1]
    print(f"START TRANSCRIPTID is {start}, END TRANSCRIPTID is {end}")
    get_person_sql(start, end, "../output", str(i+1).zfill(2))
    i+=1

```

```
##### PROCESSING BATCH 01 #####
```

```
START TRANSCRIPTID is 108.0, END TRANSCRIPTID is 273690.6
```

```
### Downloading transcript person details, start transcriptid is 108.0, end
transcriptid is 273690.6 ###
```

```

# descriptive statistics
<class 'pandas.core.frame.DataFrame'>
Index: 10071223 entries, 0 to 71222
Data columns (total 13 columns):
#   Column                                Dtype
---  -
0   transcriptid                          float64
1   transcriptcomponentid                 float64
2   componentorder                       int64
3   transcriptcomponenttypeid            int64
4   transcriptcomponenttypename          object
5   transcriptpersonid                   float64
6   transcriptpersonname                 object
7   proid                               float64
8   companyofperson                     object
9   speakertypeid                      int64

```

```

10 speakertypename      object
11 componenttextpreview   object
12 word_count             float64
dtypes: float64(5), int64(3), object(5)
memory usage: 5.2 GB
None

```

```

False      10071223
Name: count, dtype: int64

```

	transcriptid	transcriptcomponentid	componentorder	transcriptcomponenttypeid	transcriptpersonid	proid	speakertypeid
word_count							
count	10,071,223.0000	10,071,223.0000	10,071,223.0000				
	10,071,223.0000	10,071,223.0000	5,909,123.0000	10,071,223.0000			
	10,071,149.0000						
mean	108,677.1246	6,926,283.2333	43.8159				
3.8502	93,491.0579	35,462,998.6593	2.2365	108.9444			
std	74,045.5264	4,008,132.8939	34.4848				
1.3867	65,275.3401	34,495,923.0505	0.6662	300.3826			
min	109.0000	30,263.0000	0.0000				
1.0000	1.0000	1.0000	1.0000	1.0000	1.0000		
25%	43,312.0000	3,335,643.5000	17.0000				
3.0000	24,153.0000	6,611,971.0000	2.0000	14.0000			
50%	102,305.0000	7,127,760.0000	37.0000				
4.0000	102,155.0000	29,247,240.0000	2.0000	41.0000			
75%	165,285.0000	10,409,546.5000	62.0000				
4.0000	142,163.0000	49,853,710.0000	3.0000	92.0000			
max	273,690.0000	17,307,751.0000	373.0000				
8.0000	232,149.0000	272,248,561.0000	5.0000	17,178.0000			

```

# compare sample with wrds_transcript_detail
_merge

```

```

both      10168273
right_only      2157
left_only      0
Name: count, dtype: int64

```

```

# saving zip compressed pickle as wrds_transcript_person_01

```

```

##### PROCESSING BATCH 02 #####

```

```

START TRANSCRIPTID is 273690.6, END TRANSCRIPTID is 556694.6000000001

```

```

### Downloading transcript person details, start transcriptid is 273690.6, end
transcriptid is 556694.6000000001 ###

```

```

# descriptive statistics

```

```

<class 'pandas.core.frame.DataFrame'>
Index: 8907751 entries, 0 to 407750
Data columns (total 13 columns):
#   Column                                Dtype
---  -
0   transcriptid                          float64
1   transcriptcomponentid                 float64
2   componentorder                       int64
3   transcriptcomponenttypeid             int64
4   transcriptcomponenttypename           object
5   transcriptpersonid                   float64
6   transcriptpersonname                  object
7   proid                                float64
8   companyofperson                      object
9   speakertypeid                       int64
10  speakertypename                     object
11  componenttextpreview                  object
12  word_count                           float64
dtypes: float64(5), int64(3), object(5)
memory usage: 4.6 GB
None

```

```

False      8907751
Name: count, dtype: int64

```

	transcriptid	transcriptcomponentid	componentorder	transcriptcomponenttypeid	transcriptpersonid	proid	speakertypeid	word_count
count	8,907,751.0000	8,907,751.0000	8,907,751.0000	8,907,751.0000	6,650,548.0000	8,907,751.0000	8,907,531.0000	
mean	409,930.1082	19,333,919.7867	39.1050					
std	81,514.4350	3,402,542.0991	32.9075					
min	273,691.0000	13,576,724.0000	0.0000					
1.0000	1.0000	65,510.0000	1.0000					
25%	338,615.0000	16,400,088.5000	15.0000					
3.0000	98,444.0000	11,519,963.0000	2.0000					
50%	408,454.0000	19,288,359.0000	32.0000					
4.0000	148,173.0000	37,451,157.0000	2.0000					
75%	480,411.0000	22,205,567.5000	55.0000					
max	556,694.0000	25,421,632.0000	371.0000					
8.0000	281,247.0000	275,685,615.0000	5.0000					

```

# compare sample with wrds_transcript_detail
_merge

```

```
both          9078962
right_only    9
left_only     0
Name: count, dtype: int64
```

```
# saving zip compressed pickle as wrds_transcript_person_02
```

```
##### PROCESSING BATCH 03 #####
```

```
START TRANSCRIPTID is 556694.6000000001, END TRANSCRIPTID is 866928.4000000001
```

```
### Downloading transcript person details, start transcriptid is
556694.6000000001, end transcriptid is 866928.4000000001 ###
```

```
# descriptive statistics
```

```
<class 'pandas.core.frame.DataFrame'>
```

```
Index: 8728803 entries, 0 to 228802
```

```
Data columns (total 13 columns):
```

#	Column	Dtype
0	transcriptid	float64
1	transcriptcomponentid	float64
2	componentorder	int64
3	transcriptcomponenttypeid	int64
4	transcriptcomponenttypename	object
5	transcriptpersonid	float64
6	transcriptpersonname	object
7	proid	float64
8	companyofperson	object
9	speakertypeid	int64
10	speakertypename	object
11	componenttextpreview	object
12	word_count	float64

```
dtypes: float64(5), int64(3), object(5)
```

```
memory usage: 4.4 GB
```

```
None
```

```
False      8728803
```

```
Name: count, dtype: int64
```

	transcriptid	transcriptcomponentid	componentorder	transcriptcomponenttypeid	transcriptpersonid	proid	speakertypeid	word_count
count	8,728,803.0000	8,728,803.0000	8,728,803.0000	8,728,803.0000	8,728,803.0000	6,752,125.0000	8,728,803.0000	8,728,771.0000
mean	711,206.3403	31,397,261.6100	38.0145					
	3.7511	167,875.7780	92,388,216.3179	2.2579		130.1554		

std	89,396.0848	3,418,445.0026	33.3503	
1.3457	96,911.8528	88,839,099.9155	0.7254	344.9694
min	556,697.0000	25,421,633.0000	0.0000	
1.0000	1.0000	65,145.0000	1.0000	1.0000
25%	630,877.0000	28,396,850.5000	14.0000	
3.0000	102,931.0000	25,936,635.0000	2.0000	16.0000
50%	712,593.0000	31,460,326.0000	31.0000	
4.0000	178,204.0000	52,684,191.0000	2.0000	50.0000
75%	788,610.0000	34,370,749.5000	53.0000	
4.0000	255,355.0000	141,631,556.0000	3.0000	113.0000
max	866,926.0000	37,269,245.0000	501.0000	
8.0000	310,007.0000	310,959,791.0000	5.0000	14,324.0000

```
# compare sample with wrds_transcript_detail
_merge
```

```
both          9003180
```

```
right_only    1
```

```
left_only     0
```

```
Name: count, dtype: int64
```

```
# saving zip compressed pickle as wrds_transcript_person_03
```

```
##### PROCESSING BATCH 04 #####
```

```
START TRANSCRIPTID is 866928.4000000001, END TRANSCRIPTID is 1196448.6
```

```
### Downloading transcript person details, start transcriptid is
866928.4000000001, end transcriptid is 1196448.6 ###
```

```
# descriptive statistics
```

```
<class 'pandas.core.frame.DataFrame'>
```

```
Index: 8462389 entries, 0 to 462388
```

```
Data columns (total 13 columns):
```

#	Column	Dtype
0	transcriptid	float64
1	transcriptcomponentid	float64
2	componentorder	int64
3	transcriptcomponenttypeid	int64
4	transcriptcomponenttypename	object
5	transcriptpersonid	float64
6	transcriptpersonname	object
7	proid	float64
8	companyofperson	object
9	speakertypeid	int64
10	speakertypename	object
11	componenttextpreview	object
12	word_count	float64

```
dtypes: float64(5), int64(3), object(5)
memory usage: 4.3 GB
None
```

```
False      8462389
Name: count, dtype: int64
```

	transcriptid	transcriptcomponentid	componentorder	transcriptcomponenttypeid	transcriptpersonid	proid	speakertypeid
word_count							
count	8,462,389.0000	8,462,389.0000	8,462,389.0000				
	8,462,389.0000	8,462,389.0000	6,629,225.0000	8,462,389.0000			
	8,462,365.0000						
mean	1,023,994.1919	43,050,209.9365	36.2235				
3.7631	188,791.4039	128,169,203.6765	2.2486			133.7274	
std	94,429.3420	3,410,479.7557	31.2853				
1.3599	107,834.6613	114,282,784.9573	0.7279			341.0655	
min	866,929.0000	37,269,282.0000	0.0000				
1.0000	1.0000	65,510.0000	1.0000			1.0000	
25%	941,635.0000	40,098,372.0000	14.0000				
3.0000	105,904.0000	28,042,348.0000	2.0000			17.0000	
50%	1,019,744.0000	42,962,045.0000	30.0000				
4.0000	212,186.0000	84,650,574.0000	2.0000			53.0000	
75%	1,102,267.0000	45,900,676.0000	50.0000				
4.0000	288,551.0000	244,013,386.0000	3.0000			119.0000	
max	1,196,447.0000	49,147,144.0000	529.0000				
7.0000	332,212.0000	430,051,638.0000	5.0000			13,711.0000	

```
# compare sample with wrds_transcript_detail
_merge
both      8750425
left_only      0
right_only     0
Name: count, dtype: int64
```

```
# saving zip compressed pickle as wrds_transcript_person_04
```

```
##### PROCESSING BATCH 05 #####
```

```
START TRANSCRIPTID is 1196448.6, END TRANSCRIPTID is 1509145.0
```

```
### Downloading transcript person details, start transcriptid is 1196448.6, end
transcriptid is 1509145.0 ###
```

```
# descriptive statistics
<class 'pandas.core.frame.DataFrame'>
Index: 7786576 entries, 0 to 286575
Data columns (total 13 columns):
```

#	Column	Dtype
0	transcriptid	float64
1	transcriptcomponentid	float64
2	componentorder	int64
3	transcriptcomponenttypeid	int64
4	transcriptcomponenttypename	object
5	transcriptpersonid	float64
6	transcriptpersonname	object
7	proid	float64
8	companyofperson	object
9	speakertypeid	int64
10	speakertypename	object
11	componenttextpreview	object
12	word_count	float64

dtypes: float64(5), int64(3), object(5)

memory usage: 3.9 GB

None

False 7786576

Name: count, dtype: int64

	transcriptid	transcriptcomponentid	componentorder	transcriptcomponenttypeid	transcriptpersonid	proid	speakertypeid	word_count
count	7,786,576.0000	7,786,576.0000	7,786,576.0000					
	7,786,576.0000	7,786,576.0000	6,014,602.0000	7,786,576.0000				
	7,786,556.0000							
mean	1,362,041.1016	54,574,348.6982		36.2600				
	3.7160	213,387.7329	173,557,827.9114	2.2616		136.8587		
std	86,946.7447	2,993,211.3327		35.1882				
	1.3476	119,909.2858	149,982,280.5878	0.7296		350.0424		
min	1,196,451.0000	49,147,145.0000		0.0000				
	1.0000	1.0000	65,510.0000	1.0000		1.0000		
25%	1,288,192.0000	52,036,253.7500		13.0000				
	3.0000	115,922.0000	34,363,129.0000	2.0000		17.0000		
50%	1,365,898.0000	54,628,654.5000		28.0000				
	4.0000	250,189.0000	134,328,678.0000	2.0000		53.0000		
75%	1,436,607.0000	57,166,802.2500		49.0000				
	4.0000	318,725.0000	279,654,441.0000	3.0000		121.0000		
max	1,509,145.0000	68,805,185.0000		630.0000				
	7.0000	363,978.0000	575,281,285.0000	5.0000		13,664.0000		

compare sample with wrds_transcript_detail

_merge

both 8048188

right_only 12

left_only 0

```

Name: count, dtype: int64

# saving zip compressed pickle as wrds_transcript_person_05

##### PROCESSING BATCH 06 #####

START TRANSCRIPTID is 1509145.0, END TRANSCRIPTID is 1783797.6

### Downloading transcript person details, start transcriptid is 1509145.0, end
transcriptid is 1783797.6 ###

# descriptive statistics
<class 'pandas.core.frame.DataFrame'>
Index: 7966017 entries, 0 to 466016
Data columns (total 13 columns):
#   Column                                Dtype
---  -----
0   transcriptid                          float64
1   transcriptcomponentid                 float64
2   componentorder                       int64
3   transcriptcomponenttypeid             int64
4   transcriptcomponenttypename           object
5   transcriptpersonid                   float64
6   transcriptpersonname                  object
7   proid                                float64
8   companyofperson                      object
9   speakertypeid                       int64
10  speakertypename                      object
11  componenttextpreview                  object
12  word_count                            float64
dtypes: float64(5), int64(3), object(5)
memory usage: 4.0 GB
None

False    7966017
Name: count, dtype: int64

      transcriptid  transcriptcomponentid  componentorder
transcriptcomponenttypeid  transcriptpersonid      proid  speakertypeid
word_count
count  7,966,017.0000      7,966,017.0000  7,966,017.0000
7,966,017.0000      7,966,017.0000  6,224,696.0000  7,966,017.0000
7,965,999.0000
mean  1,646,999.0383      64,826,679.9700      37.3268
3.7360      231,305.2494  224,923,399.5414      2.2662      136.5494
std    79,782.4716      2,945,669.9846      36.0036
1.3391      128,100.7275  190,462,356.2133      0.7429      347.1876
min    1,509,148.0000      59,692,724.0000      0.0000

```

1.0000	1.0000	65,510.0000	1.0000	1.0000
25%	1,579,843.0000	62,332,254.0000	13.0000	
3.0000	130,981.0000	43,211,086.0000	2.0000	16.0000
50%	1,645,669.0000	64,855,601.0000	28.0000	
4.0000	280,925.0000	222,638,290.0000	2.0000	52.0000
75%	1,717,180.0000	67,402,758.0000	50.0000	
4.0000	338,379.0000	330,075,607.0000	3.0000	122.0000
max	1,783,797.0000	69,851,159.0000	429.0000	
7.0000	396,324.0000	631,393,004.0000	5.0000	21,091.0000

```
# compare sample with wrds_transcript_detail
_merge
both      8184717
right_only 17
left_only  0
Name: count, dtype: int64
```

```
# saving zip compressed pickle as wrds_transcript_person_06
```

```
##### PROCESSING BATCH 07 #####
```

```
START TRANSCRIPTID is 1783797.6, END TRANSCRIPTID is 2038333.4000000004
```

```
### Downloading transcript person details, start transcriptid is 1783797.6, end
transcriptid is 2038333.4000000004 ###
```

```
# descriptive statistics
<class 'pandas.core.frame.DataFrame'>
Index: 7636382 entries, 0 to 136381
Data columns (total 13 columns):
#   Column                                Dtype
---  -----
0   transcriptid                          float64
1   transcriptcomponentid                 float64
2   componentorder                       int64
3   transcriptcomponenttypeid             int64
4   transcriptcomponenttypename           object
5   transcriptpersonid                   float64
6   transcriptpersonname                  object
7   proid                                float64
8   companyofperson                      object
9   speakertypeid                       int64
10  speakertypename                     object
11  componenttextpreview                  object
12  word_count                           float64
dtypes: float64(5), int64(3), object(5)
memory usage: 3.9 GB
None
```

```
False    7636382
Name: count, dtype: int64
```

```

transcriptid transcriptcomponentid componentorder
transcriptcomponenttypeid transcriptpersonid      proid speakertypeid
word_count
count 7,636,382.0000      7,636,382.0000 7,636,382.0000
7,636,382.0000      7,636,382.0000 5,900,548.0000 7,636,382.0000
7,636,375.0000
mean 1,905,059.4105      74,662,804.4320      39.0243
3.7832      252,985.4713 273,754,318.3190      2.2571      138.9140
std      70,910.1679      2,823,770.6962      57.2729
1.3429      140,387.3024 218,868,489.5263      0.7632      341.3793
min 1,783,798.0000      69,851,160.0000      0.0000
1.0000      1.0000      65,510.0000      1.0000      1.0000
25% 1,846,143.0000      72,246,059.2500      13.0000
3.0000      149,044.0000 51,730,322.0000      2.0000      16.0000
50% 1,901,171.0000      74,587,820.5000      28.0000
4.0000      304,354.0000 253,852,615.0000      2.0000      53.0000
75% 1,959,919.0000      77,001,602.7500      49.0000
4.0000      365,037.0000 531,574,041.0000      3.0000      126.0000
max 2,038,333.0000      81,257,025.0000      1,867.0000
7.0000      444,852.0000 677,194,656.0000      5.0000      19,609.0000

```

```

# compare sample with wrds_transcript_detail
_merge
both      7793249
right_only      10
left_only      0
Name: count, dtype: int64

```

```
# saving zip compressed pickle as wrds_transcript_person_07
```

```
##### PROCESSING BATCH 08 #####
```

```
START TRANSCRIPTID is 2038333.4000000004, END TRANSCRIPTID is 2391958.8000000003
```

```
### Downloading transcript person details, start transcriptid is
2038333.4000000004, end transcriptid is 2391958.8000000003 ###
```

```

# descriptive statistics
<class 'pandas.core.frame.DataFrame'>
Index: 7388901 entries, 0 to 388900
Data columns (total 13 columns):
#   Column      Dtype
---  -----  ---
0   transcriptid float64

```

```

1 transcriptcomponentid float64
2 componentorder int64
3 transcriptcomponenttypeid int64
4 transcriptcomponenttypename object
5 transcriptpersonid float64
6 transcriptpersonname object
7 proid float64
8 companyofperson object
9 speakertypeid int64
10 speakertypename object
11 componenttextpreview object
12 word_count float64
dtypes: float64(5), int64(3), object(5)
memory usage: 3.8 GB
None

False 7388901
Name: count, dtype: int64

transcriptid transcriptcomponentid componentorder
transcriptcomponenttypeid transcriptpersonid proid speakertypeid
word_count
count 7,388,901.0000 7,388,901.0000 7,388,901.0000
7,388,901.0000 7,388,901.0000 5,704,017.0000 7,388,901.0000
7,388,883.0000
mean 2,194,897.1724 85,449,688.3181 35.6513
3.7989 278,421.6882 320,548,928.2074 2.2470 144.2129
std 104,478.0932 3,251,300.9701 35.1884
1.3505 154,260.4297 245,016,099.9314 0.7704 326.1269
min 2,038,335.0000 79,803,979.0000 0.0000
1.0000 1.0000 66,522.0000 1.0000 1.0000
25% 2,102,915.0000 82,660,030.0000 12.0000
3.0000 171,132.0000 60,838,314.0000 2.0000 17.0000
50% 2,177,597.0000 85,328,439.0000 26.0000
4.0000 327,725.0000 280,319,821.0000 2.0000 58.0000
75% 2,278,295.0000 88,202,103.0000 47.0000
4.0000 398,325.0000 575,130,116.0000 3.0000 138.0000
max 2,391,954.0000 91,132,697.0000 476.0000
7.0000 502,958.0000 1,680,487,720.0000 5.0000 17,752.0000

# compare sample with wrds_transcript_detail
_merge
both 7548768
right_only 53
left_only 0
Name: count, dtype: int64

# saving zip compressed pickle as wrds_transcript_person_08

```

PROCESSING BATCH 09

START TRANSCRIPTID is 2391958.8000000003, END TRANSCRIPTID is 2736943.6000000006

Downloading transcript person details, start transcriptid is 2391958.8000000003, end transcriptid is 2736943.6000000006

descriptive statistics

<class 'pandas.core.frame.DataFrame'>

Index: 7552993 entries, 0 to 52992

Data columns (total 13 columns):

#	Column	Dtype
0	transcriptid	float64
1	transcriptcomponentid	float64
2	componentorder	int64
3	transcriptcomponenttypeid	int64
4	transcriptcomponenttypename	object
5	transcriptpersonid	float64
6	transcriptpersonname	object
7	proid	float64
8	companyofperson	object
9	speakertypeid	int64
10	speakertypename	object
11	componenttextpreview	object
12	word_count	float64

dtypes: float64(5), int64(3), object(5)

memory usage: 3.9 GB

None

False 7552993

Name: count, dtype: int64

	transcriptid	transcriptcomponentid	componentorder	transcriptcomponenttypeid	transcriptpersonid	proid	speakertypeid	word_count
count	7,552,993.0000	7,552,993.0000	7,552,993.0000					
	7,552,993.0000	7,552,993.0000	5,532,666.0000	7,552,993.0000				
	7,552,897.0000							
mean	2,562,626.2289	96,840,440.6516	35.5153					
3.7855	302,793.5584	455,271,453.7501	2.2509				144.4976	
std	98,505.1535	3,314,070.5015	37.3991					
1.3261	177,177.1011	446,482,774.9332	0.7519				335.0350	
min	2,391,960.0000	91,132,698.0000	0.0000					
1.0000	1.0000	65,043.0000	1.0000				1.0000	
25%	2,479,859.0000	93,999,459.0000	12.0000					
3.0000	174,024.0000	108,759,599.0000	2.0000				18.0000	

```

50%    2,559,080.0000    96,778,538.0000    26.0000
4.0000    340,253.0000    331,299,081.0000    2.0000    58.0000
75%    2,648,938.0000    99,746,288.0000    47.0000
4.0000    449,145.0000    636,724,090.0000    3.0000    136.0000
max    2,736,942.0000    102,590,129.0000    1,182.0000
7.0000    566,407.0000    1,825,174,193.0000    5.0000    19,600.0000

```

```

# compare sample with wrds_transcript_detail
_merge

```

```

both          7682025
right_only    18
left_only     0
Name: count, dtype: int64

```

```

# saving zip compressed pickle as wrds_transcript_person_09

```

```

##### PROCESSING BATCH 10 #####

```

```

START TRANSCRIPTID is 2736943.6000000006, END TRANSCRIPTID is 3175090.0

```

```

### Downloading transcript person details, start transcriptid is
2736943.6000000006, end transcriptid is 3175090.0 ###

```

```

# descriptive statistics

```

```

<class 'pandas.core.frame.DataFrame'>

```

```

Index: 7223709 entries, 0 to 223708

```

```

Data columns (total 13 columns):

```

#	Column	Dtype
0	transcriptid	float64
1	transcriptcomponentid	float64
2	componentorder	int64
3	transcriptcomponenttypeid	int64
4	transcriptcomponenttypename	object
5	transcriptpersonid	float64
6	transcriptpersonname	object
7	proid	float64
8	companyofperson	object
9	speakertypeid	int64
10	speakertypename	object
11	componenttextpreview	object
12	word_count	float64

```

dtypes: float64(5), int64(3), object(5)

```

```

memory usage: 3.7 GB

```

```

None

```

```

False    7223709

```

```

Name: count, dtype: int64

```

	transcriptid	transcriptcomponentid	componentorder		
	transcriptcomponenttypeid	transcriptpersonid		proid	speakertypeid
word_count					
count	7,223,709.0000	7,223,709.0000	7,223,709.0000		
	7,223,709.0000	7,223,709.0000	5,318,675.0000	7,223,709.0000	
	7,223,683.0000				
mean	2,942,774.5061	108,185,329.4372	34.2042		
	3.7941	334,719.7189	637,652,266.7062	2.2499	142.9803
std	124,205.6460	3,190,862.9980	33.9712		
	1.3421	197,262.4136	616,362,430.5450	0.7644	326.1436
min	2,736,946.0000	102,590,354.0000	0.0000		
	1.0000	1.0000	66,522.0000	1.0000	1.0000
25%	2,833,692.0000	105,466,410.0000	12.0000		
	3.0000	182,536.0000	142,082,611.0000	2.0000	19.0000
50%	2,937,578.0000	108,279,752.0000	25.0000		
	4.0000	365,232.0000	537,345,353.0000	2.0000	57.0000
75%	3,044,701.0000	110,950,768.0000	45.0000		
	4.0000	505,743.0000	700,048,953.0000	3.0000	135.0000
max	3,175,090.0000	113,613,786.0000	785.0000		
	7.0000	623,418.0000	1,885,828,003.0000	5.0000	24,325.0000

```
# compare sample with wrds_transcript_detail
_merge
both          7322416
right_only    3
left_only     0
Name: count, dtype: int64
```

```
# saving zip compressed pickle as wrds_transcript_person_10
```

```
[47]: del transcripts
```

Check that wrds_transcript_person files have data for all transcripts

```
[48]: import glob
from functools import reduce
filelist = glob.glob(os.path.abspath("../output/")+"\\wrds_transcript_person_*")
# filelist.append(filelist.pop(0))
filelist
```

```
[48]: ['c:\\Users\\flakej\\Dropbox\\Research\\Data\\CIQ_Transcripts\\2024\\output\\wrds_transcript_person_01.pkl',
       'c:\\Users\\flakej\\Dropbox\\Research\\Data\\CIQ_Transcripts\\2024\\output\\wrds_transcript_person_02.pkl',
       'c:\\Users\\flakej\\Dropbox\\Research\\Data\\CIQ_Transcripts\\2024\\output\\wrds_transcript_person_03.pkl',
       'c:\\Users\\flakej\\Dropbox\\Research\\Data\\CIQ_Transcripts\\2024\\output\\wrds_transcript_person_04.pkl']
```

```
s_transcript_person_04.pkl',
'c:\\Users\\flakej\\Dropbox\\Research\\Data\\CIQ_Transcripts\\2024\\output\\wrd
s_transcript_person_05.pkl',
'c:\\Users\\flakej\\Dropbox\\Research\\Data\\CIQ_Transcripts\\2024\\output\\wrd
s_transcript_person_06.pkl',
'c:\\Users\\flakej\\Dropbox\\Research\\Data\\CIQ_Transcripts\\2024\\output\\wrd
s_transcript_person_07.pkl',
'c:\\Users\\flakej\\Dropbox\\Research\\Data\\CIQ_Transcripts\\2024\\output\\wrd
s_transcript_person_08.pkl',
'c:\\Users\\flakej\\Dropbox\\Research\\Data\\CIQ_Transcripts\\2024\\output\\wrd
s_transcript_person_09.pkl',
'c:\\Users\\flakej\\Dropbox\\Research\\Data\\CIQ_Transcripts\\2024\\output\\wrd
s_transcript_person_10.pkl']
```

```
[60]: def get_component(filelist):
        df_out = pd.DataFrame()
        for file in filelist:
            print(f"\n### Processing {file} ###")
            df1 = pd.read_pickle(file, compression="zip")
            df1['saved'] = 1
            df1 = df1[['transcriptid', 'saved']].drop_duplicates()
            print(f"Input DF contains {df1.shape[0]} obs")
            df_out = pd.concat([df_out, df1], ignore_index=True)
            print(f"Aggregated DF contains {df_out.shape[0]} obs")
        return df_out
```

```
[61]: df1.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 10071223 entries, 0 to 71222
Data columns (total 13 columns):
 #   Column                                Dtype
---  -
 0   transcriptid                          float64
 1   transcriptcomponentid                 float64
 2   componentorder                       int64
 3   transcriptcomponenttypeid            int64
 4   transcriptcomponenttypename          object
 5   transcriptpersonid                   float64
 6   transcriptpersonname                  object
 7   proid                                float64
 8   companyofperson                      object
 9   speakertypeid                       int64
10   speakertypename                     object
11   componenttextpreview                 object
12   word_count                           float64
dtypes: float64(5), int64(3), object(5)
memory usage: 5.2 GB
```

```
[62]: df1.tail()
```

```
[62]:      transcriptid transcriptcomponentid componentorder
transcriptcomponenttypeid transcriptcomponenttypename
transcriptpersonid transcriptpersonname proid companyofperson
speakertypeid speakertypename \
71218 273,690.0000 13,576,719.0000 51
4 Answer 148,453.0000 Constantine
Karayannopoulos 26,874,765.0000 None 2 Executives
71219 273,690.0000 13,576,720.0000 52
4 Answer 171,086.0000
Mark Smith 102,220,392.0000 None 2 Executives
71220 273,690.0000 13,576,721.0000 53
7 Question and Answer Operator Message 1.0000
Operator NaN None 1 Operator
71221 273,690.0000 13,576,722.0000 54
4 Answer 171,086.0000
Mark Smith 102,220,392.0000 None 2 Executives
71222 273,690.0000 13,576,723.0000 55
7 Question and Answer Operator Message 1.0000
Operator NaN None 1 Operator

      componenttextpreview word_count
71218 Well, again as much as I like to agree with yo... 78.0000
71219 Thank you, Constantine. 3.0000
71220 This is all the time you for questions. I woul... 25.0000
71221 Okay. Thanks, operator. I'd like to thank ever... 172.0000
71222 Ladies and gentlemen that concludes today's co... 21.0000
```

```
[63]: df = get_component(filelist)
df.info()
```

```
### Processing c:\Users\flakej\Dropbox\Research\Data\CIQ_Transcripts\2024\output
\wrds_transcript_person_01.pkl ###
```

```
Input DF contains 152179 obs
```

```
Aggregated DF contains 152179 obs
```

```
### Processing c:\Users\flakej\Dropbox\Research\Data\CIQ_Transcripts\2024\output
\wrds_transcript_person_02.pkl ###
```

```
Input DF contains 154327 obs
```

```
Aggregated DF contains 306506 obs
```

```
### Processing c:\Users\flakej\Dropbox\Research\Data\CIQ_Transcripts\2024\output
\wrds_transcript_person_03.pkl ###
```

```
Input DF contains 154334 obs
```

```
Aggregated DF contains 460840 obs
```

```

### Processing c:\Users\flakej\Dropbox\Research\Data\CIQ_Transcripts\2024\output
\wrds_transcript_person_04.pkl ###
Input DF contains 154336 obs
Aggregated DF contains 615176 obs

### Processing c:\Users\flakej\Dropbox\Research\Data\CIQ_Transcripts\2024\output
\wrds_transcript_person_05.pkl ###
Input DF contains 154324 obs
Aggregated DF contains 769500 obs

### Processing c:\Users\flakej\Dropbox\Research\Data\CIQ_Transcripts\2024\output
\wrds_transcript_person_06.pkl ###
Input DF contains 154318 obs
Aggregated DF contains 923818 obs

### Processing c:\Users\flakej\Dropbox\Research\Data\CIQ_Transcripts\2024\output
\wrds_transcript_person_07.pkl ###
Input DF contains 154328 obs
Aggregated DF contains 1078146 obs

### Processing c:\Users\flakej\Dropbox\Research\Data\CIQ_Transcripts\2024\output
\wrds_transcript_person_08.pkl ###
Input DF contains 154284 obs
Aggregated DF contains 1232430 obs

### Processing c:\Users\flakej\Dropbox\Research\Data\CIQ_Transcripts\2024\output
\wrds_transcript_person_09.pkl ###
Input DF contains 154318 obs
Aggregated DF contains 1386748 obs

### Processing c:\Users\flakej\Dropbox\Research\Data\CIQ_Transcripts\2024\output
\wrds_transcript_person_10.pkl ###
Input DF contains 154333 obs
Aggregated DF contains 1541081 obs
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1541081 entries, 0 to 1541080
Data columns (total 2 columns):
#   Column          Non-Null Count  Dtype
---  -
0   transcriptid     1541081 non-null  float64
1   saved            1541081 non-null  int64
dtypes: float64(1), int64(1)
memory usage: 23.5 MB

```

```
[64]: df.duplicated().value_counts()
```

```
[64]: False      1541081
      Name: count, dtype: int64
```

Compare with wrds_transcript_detail

```
[65]: wrds_transcript_detail.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 1575627 entries, 0 to 75626
Data columns (total 16 columns):
 #   Column                                  Non-Null Count  Dtype
---  -
 0   companyid                             1575627 non-null object
 1   keydevid                               1575627 non-null object
 2   transcriptid                           1575627 non-null object
 3   headline                               1575627 non-null object
 4   mostimportantdateutc                   1575627 non-null object
 5   mostimportanttimeutc                   1575627 non-null object
 6   keydeveventtypeid                     1575627 non-null object
 7   keydeveventtypename                   1575627 non-null object
 8   companyname                           1574753 non-null object
 9   transcriptcollectiontypeid             1575627 non-null int64
10   transcriptcollectiontypename           1575627 non-null object
11   transcriptpresentationtypeid           1575627 non-null int64
12   transcriptpresentationtypename         1575627 non-null object
13   transcriptcreationdate_utc             1575627 non-null object
14   transcriptcreationtime_utc             1575627 non-null object
15   audiolengthsec                         1532987 non-null object
dtypes: int64(2), object(14)
memory usage: 1.7 GB
```

```
[67]: df2 = df.merge(right=wrds_transcript_detail, on='transcriptid', how="outer",
      ↪indicator=True)
```

The previous step shows 2,281 right_only records, inspect them

```
[68]: df2._merge.value_counts()
```

```
[68]: _merge
both      1573346
right_only    2281
left_only      0
Name: count, dtype: int64
```

```
[75]: len(df2[df2._merge=='right_only'].sort_values(by='transcriptid'))
```

```
[75]: 2281
```

```
[99]: right_only = df2[df2._merge=='right_only'].sort_values(by='transcriptid')
```

```
[100]: trans_r_only = right_only['transcriptid'].astype(int).unique().tolist()
len(trans_r_only)
```

[100]: 2276

Download missing records

```
[80]: import wrds
db = wrds.Connection(wrds_username = 'j4ffle')
```

Loading library list...

Done

```
[81]: new = db.raw_sql("select distinct * from ciq_transcripts.wrds_transcript_person_
↳where transcriptid in "+str(tuple(trans_r_only)))
```

```
[108]: new_det = db.raw_sql("select distinct * from ciq_transcripts.
↳wrds_transcript_detail where transcriptid in "+str(tuple(trans_r_only)))
new_det.head()
```

```
[108]:      companyid      keydevid transcriptid
headline mostimportantdateutc mostimportanttimeutc keydeveventypeid
keydeveventtypename      companyname \
0   873,976.0000 137,253,286.0000 151,041.0000  BNP Paribas, Q2 2011 Earnings
Call, Aug 02, 2011      2011-08-02      13:30:00      48.0000
Earnings Calls      BNP Paribas SA
1   29,279.0000  6,180,480.0000  16,139.0000  Harman International Industries
Inc., Q2 2009 ...      2009-02-04      21:40:00      48.0000
Earnings Calls Harman International Industries, Incorporated
2   36,118.0000  5,979,678.0000  15,106.0000  Geeknet, Inc., Q1 2009 Earnings
Call, Nov-25-2008      2008-11-25      22:00:00      48.0000
Earnings Calls      Geeknet, Inc.
3   354,995.0000 118,231,910.0000  96,055.0000  International Speedway Corp.,
Q4 2010 Earnings...      2011-01-27      14:00:00      48.0000
Earnings Calls      International Speedway Corporation
4  4,169,095.0000  5,606,560.0000  11,510.0000  Triple-S Management
Corporation, Q2 2008 Earni...      2008-08-05      14:00:00
48.0000      Earnings Calls      Triple-S Management Corporation

      transcriptcollectiontypeid transcriptcollectiontypename
transcriptpresentationtypeid transcriptpresentationtypename
transcriptcreationdate_utc transcriptcreationtime_utc audiolengthsec
0      6      SA Edited Copy
5      Final      2011-08-03
15:02:52      6,533.0000
1      6      SA Edited Copy
5      Final      2009-02-05
14:44:54      NaN
```

2		6	SA Edited Copy
5		Final	2008-11-26
05:08:05	NaN		
3		6	SA Edited Copy
5		Final	2011-01-27
19:25:21	3,171.0000		
4		6	SA Edited Copy
5		Final	2008-08-22
21:59:16	NaN		

```
[113]: df[df.transcriptid.isin([166319,151041])]
```

```
[113]:      transcriptid  saved
109047  166,319.0000      1
```

```
[115]: db.raw_sql("select distinct * from ciq_transcripts.wrds_transcript_detail where
↳transcriptid in "+str(tuple([166319,151041])))
```

```
[115]:      companyid      keydevid transcriptid
headline mostimportantdateutc mostimportanttimeutc keydeveventtypeid
keydeveventtypename      companyname transcriptcollectiontypeid \
0 873,976.0000 137,253,286.0000 151,041.0000 BNP Paribas, Q2 2011 Earnings
Call, Aug 02, 2011      2011-08-02      13:30:00      48.0000
Earnings Calls BNP Paribas SA      6
1 873,976.0000 137,253,286.0000 166,319.0000 BNP Paribas, Q2 2011 Earnings
Call, Aug 02, 2011      2011-08-02      13:30:00      48.0000
Earnings Calls BNP Paribas SA      8
```

	transcriptcollectiontypename	transcriptpresentationtypeid	transcriptpresentationtypename	transcriptcreationdate_utc	transcriptcreationtime_utc	audiolengthsec
0	SA Edited Copy					5
Final		2011-08-03		15:02:52		6,533.0000
1	Audited Copy					5
Final		2011-09-08		07:34:34		6,533.0000

Same transcript information, but the first id = 151041 is not in the person or full-transcript data set

```
[92]: new = new.merge(right=df2.loc[df2._merge=='right_only',['transcriptid']],
↳how="outer", indicator=True)
```

```
[93]: new._merge.value_counts()
```

```
[93]: _merge
right_only      2280
both              79
left_only         0
```

Name: count, dtype: int64

Of the 2,280 transcripts from detail without corresponding, only one of them has full or person transcript data

```
[94]: new = new.drop(columns=['_merge'])
      new.head()
```

```
[94]: transcriptid transcriptcomponentid componentorder transcriptcomponenttypeid
transcriptcomponenttypename transcriptpersonid transcriptpersonname proid
companyofperson speakertypeid speakertypeename \
0      108.0000      30,233.0000      50.0000      3.0000
Question      684.0000      Ajit Pai      NaN      Thomas Weisel Partners
3.0000      Analysts
1      108.0000      30,198.0000      15.0000      3.0000
Question      932.0000      Deane Dray      NaN      Goldman Sachs
3.0000      Analysts
2      108.0000      30,202.0000      19.0000      3.0000
Question      896.0000      Darryl Pardi      NaN      Merrill Lynch
3.0000      Analysts
3      108.0000      30,244.0000      61.0000      4.0000
Answer      12.0000      Adrian Dillon      NaN      None
2.0000      Executives
4      108.0000      30,186.0000      3.0000      2.0000
Presenter Speech      650.0000      William P. Sullivan      NaN
None      2.0000      Executives

      componenttextpreview      word_count
0      Got it. Okay, thank you so much.      7.0000
1      So, this, but there are no other changes in ha...      33.0000
2      Hey, good evening guys.      4.0000
3      I can't do it off the top of my head.      11.0000
4      Thanks, Hilliard and hello everyone. We are p...      776.0000
```

```
[95]: new.to_pickle("../output/wrds_transcript_person_x.pkl", compression='zip')
```