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

In [2]:
```
# read the csv into jupyter notebook
projectdata0 = pd.read_csv("/Users/kate/Downloads/14100036-eng/14100036.csv")
```

In [4]:
```
# Get distinct value of Actual hours worked and select 'Total actual hours (main
projectdata0['Actual hours worked'].unique()
```

Out[4]:
```
array(['Total employed, all hours', '0 hours', '1 to 14 hours',
       '15 to 29 hours', '30 to 34 hours', '35 to 39 hours', '40 hours',
       '41 to 49 hours', '50 hours or more',
       'Total actual hours (main job)',
       'Average actual hours (all workers, main job)',
       'Average actual hours (worked in reference week, main job)'],
      dtype=object)
```

In [5]:
```
# Since there are hirearchy in NAICS. Select detail industry only, drop all indu
# and Services-producing sector,
projectdata0['North American Industry Classification System (NAICS)'].unique()
```

Out[5]:
```
array(['Total employed, all industries', 'Goods-producing sector',
       'Agriculture [111-112, 1100, 1151-1152]',
       'Forestry, fishing, mining, quarrying, oil and gas [21, 113-114, 1153, 21
00]',
       'Utilities [22]', 'Construction [23]', 'Manufacturing [31-33]',
       'Services-producing sector',
       'Wholesale and retail trade [41, 44-45]',
       'Transportation and warehousing [48-49]',
       'Finance, insurance, real estate, rental and leasing [52-53]',
       'Professional, scientific and technical services [54]',
       'Business, building and other support services [55-56]',
       'Educational services [61]',
       'Health care and social assistance [62]',
       'Information, culture and recreation [51, 71]',
       'Accommodation and food services [72]',
       'Other services (except public administration) [81]',
       'Public administration [91]'], dtype=object)
```

In [6]:
```
# Get distinct value of Class of worker
projectdata0['Class of worker'].unique()
```

Out[6]:
```
array(['Total employed', 'Employees', 'Self-employed'], dtype=object)
```

In [48]:
```
#Convert date format
projectdata0['YEARMM']=pd.to_datetime(projectdata0['REF_DATE'])

# Drop reduntant columns
drop_columns=['REF_DATE','DGUID','UOM_ID','SCALAR_FACTOR','SCALAR_ID','VECTOR','
projectdata = projectdata0.drop(drop_columns,axis=1)

# Define NAICS
NAICS = ['Agriculture [111-112, 1100, 1151-1152]',
       'Forestry, fishing, mining, quarrying, oil and gas [21, 113-114, 1153, 21
       'Utilities [22]', 'Construction [23]', 'Manufacturing [31-33]',
       'Wholesale and retail trade [41, 44-45]',
       'Transportation and warehousing [48-49]',
       'Finance, insurance, real estate, rental and leasing [52-53]',
       'Professional, scientific and technical services [54]',
       'Business, building and other support services [55-56]',
```

```
        'Educational services [61]',
        'Health care and social assistance [62]',
        'Information, culture and recreation [51, 71]',
        'Accommodation and food services [72]',
        'Other services (except public administration) [81]',
        'Public administration [91]']

# Clean data and get value total actual hours for both employee and self-employe
ts = projectdata[(projectdata['North American Industry Classification System (NA
                 &(projectdata['Actual hours worked']=='Total actual hours (main
                & (projectdata['Class of worker']=='Total employed')]
df_wh_new = ts[(ts['YEARMM']>='20000101')]
df_wh_f = df_wh_new.drop(['Actual hours worked','Class of worker'],axis=1)
display(df_wh_f.head())
```

	GEO	North American Industry Classification System (NAICS)	Sex	VALUE	YEARMM
3479889	Canada	Agriculture [111-112, 1100, 1151-1152]	Both sexes	12068.4	2000-01-01
3479890	Canada	Agriculture [111-112, 1100, 1151-1152]	Males	9336.9	2000-01-01
3479891	Canada	Agriculture [111-112, 1100, 1151-1152]	Females	2731.5	2000-01-01
3479892	Canada	Forestry, fishing, mining, quarrying, oil and ...	Both sexes	9904.2	2000-01-01
3479893	Canada	Forestry, fishing, mining, quarrying, oil and ...	Males	8769.0	2000-01-01

In [49]:
```
# Write data out
df_wh_f.to_csv (r'/Users/kate/Downloads/603_workhours.csv', index = False, heade
```

In [ ]:
```

```

SQL Queries

SELECT * FROM `bowen.li`.working_hour_data;

reference_date	geo	naics	sex	work_value
2000-01-00 00:00:00	Canada	Agriculture [111-112, 1100, 1151-1152]	Both sexes	12068
2000-01-00 00:00:00	Canada	Agriculture [111-112, 1100, 1151-1152]	Males	9337
2000-01-00 00:00:00	Canada	Agriculture [111-112, 1100, 1151-1152]	Females	2732
2000-01-00 00:00:00	Canada	Forestry, fishing, mining, quarrying, oi	Both sexes	9904
2000-01-00 00:00:00	Canada	Forestry, fishing, mining, quarrying, oi	Males	8769
2000-01-00 00:00:00	Canada	Forestry, fishing, mining, quarrying, oi	Females	1135
2000-01-00 00:00:00	Canada	Utilities [22]	Both sexes	4172
2000-01-00 00:00:00	Canada	Utilities [22]	Males	3335
2000-01-00 00:00:00	Canada	Utilities [22]	Females	837
2000-01-00 00:00:00	Canada	Construction [23]	Both sexes	25023
2000-01-00 00:00:00	Canada	Construction [23]	Males	22821
2000-01-00 00:00:00	Canada	Construction [23]	Females	2203
2000-01-00 00:00:00	Canada	Manufacturing [31-33]	Both sexes	84117
2000-01-00 00:00:00	Canada	Manufacturing [31-33]	Males	62027
2000-01-00 00:00:00	Canada	Manufacturing [31-33]	Females	22090
2000-01-00 00:00:00	Canada	Wholesale and retail trade [41, 44-45]	Both sexes	72300
2000-01-00 00:00:00	Canada	Wholesale and retail trade [41, 44-45]	Males	42719
2000-01-00 00:00:00	Canada	Wholesale and retail trade [41, 44-45]	Females	29581
2000-01-00 00:00:00	Canada	Transportation and warehousing [48-...	Both sexes	28697
2000-01-00 00:00:00	Canada	Transportation and warehousing [48-...	Males	23405
2000-01-00 00:00:00	Canada	Transportation and warehousing [48-...	Females	5292
2000-01-00 00:00:00	Canada	Finance, insurance, real estate, rental	Both sexes	29289

-- Return a list of avg working hour for the last 20 years

```
select reference_date, avg(work_value) from working_hour_data
where geo = 'Canada'
and sex = 'Both sexes'
group by year(reference_date);
```

reference_date	avg(work_value)
2000-01-00 00:00:00	31785.7760
2001-01-00 00:00:00	31518.8802
2002-01-00 00:00:00	32020.4063
2003-01-00 00:00:00	32225.8750
2004-01-00 00:00:00	33071.4896
2005-01-00 00:00:00	33748.0469
2006-01-00 00:00:00	34063.8646
2007-01-00 00:00:00	34896.9792
2008-01-00 00:00:00	35113.7500
2009-01-00 00:00:00	33657.6094
2010-01-00 00:00:00	34574.7188
2011-01-00 00:00:00	35257.7292
2012-01-00 00:00:00	35921.3698
2013-01-00 00:00:00	36236.2448
2014-01-00 00:00:00	35960.1146
2015-01-00 00:00:00	36574.6927
2016-01-00 00:00:00	36847.7500
2017-01-00 00:00:00	37329.3125
2018-01-00 00:00:00	38259.8125
2019-01-00 00:00:00	38201.9375
2020-01-00 00:00:00	34360.4250

-- Return the total workhour in 2019 and 2020 for different industries.

```
select d19.naics,avg(d19.work_value) as 19_work_hour, avg(d20.work_value) as
d20_work_hour
from working_hour_data d19, working_hour_data d20
where YEAR(d19.reference_date)=2019 and year(d20.reference_date)=2020
and d19.sex = 'Both sexes' and d20.sex = 'Both sexes'
and d19.geo = 'Canada' and d20.geo = 'Canada'
and d19.naics=d20.naics
group by naics;
```

naics	19_work_hour	d20_work_hour
Accommodation and food services [72]	33257.0000	22510.3000
Agriculture [111-112, 1100, 1151-1152]	11796.6667	11426.3000
Business, building and other support ser	23408.4167	20186.0000
Construction [23]	52954.7500	46128.5000
Educational services [61]	37565.5000	33310.3000
Finance, insurance, real estate, rental	39786.3333	39571.3000
Forestry, fishing, mining, quarrying, oi	13320.1667	11770.6000
Health care and social assistance [62]	72951.5833	68348.9000
Information, culture and recreation [51,	22888.9167	18991.7000
Manufacturing [31-33]	62873.1667	58213.0000
Other services (except public administra	25917.5000	21013.6000
Professional, scientific and technical s	52558.2500	50643.6000
Public administration [91]	32246.6667	32047.3000
Transportation and warehousing [48-49]	37641.3333	31848.4000
Utilities [22]	4921.6667	4954.6000
Wholesale and retail trade [41, 44-45]	87143.0833	78802.4000

-- Return the maximum working hour across Canada for each industry in 2020.

```
SELECT A.NAICS,MAX(A.WV) FROM (
select naics,avg(work_value) WV from working_hour_data
where  sex = 'Both sexes'
and year(reference_date)=2020
group by geo,naics) A GROUP BY  A.NAICS
order by 2 desc;
```

NAICS	MAX(A.WV)
Wholesale and retail trade [41, 44-45]	78802.4000
Health care and social assistance [62]	68348.9000
Manufacturing [31-33]	58213.0000
Professional, scientific and technical s	50643.6000
Construction [23]	46128.5000
Finance, insurance, real estate, rental	39571.3000
Educational services [61]	33310.3000
Public administration [91]	32047.3000
Transportation and warehousing [48-49]	31848.4000
Accommodation and food services [72]	22510.3000
Other services (except public administra	21013.6000
Business, building and other support ser	20186.0000
Information, culture and recreation [51,	18991.7000
Forestry, fishing, mining, quarrying, oi	11770.6000
Agriculture [111-112, 1100, 1151-1152]	11426.3000
Utilities [22]	4954.6000

-- Return a list that max hour value for each industry and the province this value belong.

```
SELECT pwh.geo, pwh.naics,pwh.wv
FROM (SELECT A.naics,max(wv) mwv FROM (
select naics,avg(work_value) AS WV from working_hour_data
where  sex = 'Both sexes'
and geo <> 'Canada'
and year(reference_date)=2020
group by geo,naics) A GROUP BY a.naics) mw,
(select geo, naics,avg(work_value) wv from working_hour_data
where  sex = 'Both sexes'
and geo <>'Canada'
and year(reference_date)=2020
group by geo,naics) pwh WHERE mw.naics=pwh.naics and mw.mwv=pwh.wv;
```

geo	naics	wv
Ontario	Accommodation and food services [72]	7749.4000
Ontario	Agriculture [111-112, 1100, 1151-1152]	2769.9000
Ontario	Business, building and other support ser	8627.3000
Ontario	Construction [23]	17106.2000
Ontario	Educational services [61]	12834.3000
Ontario	Finance, insurance, real estate, rental	19897.3000
Alberta	Forestry, fishing, mining, quarrying, oi	5283.7000
Ontario	Health care and social assistance [62]	24933.3000
Ontario	Information, culture and recreation [51,	7782.9000
Ontario	Manufacturing [31-33]	25858.0000
Ontario	Other services (except public administra	7766.8000
Ontario	Professional, scientific and technical s	22511.5000
Ontario	Public administration [91]	12742.8000
Ontario	Transportation and warehousing [48-49]	11931.7000
Ontario	Utilities [22]	1909.2000
Ontario	Wholesale and retail trade [41, 44-45]	29888.2000

-- Return the larger total hour for different industries for male and female.

```
select whm.naics,whm.mwv as male_wh, whf.mwv as female_wh
from (select A.naics, A.sex,max(WV) mwv from(
select sex,naics,avg(work_value) AS WV from working_hour_data
where  sex = 'Males'
and geo <> 'Canada'
and year(reference_date)=2020
group by sex,naics) A group by naics,sex) whm,
(select A.naics, A.sex,max(WV) mwv from(
select sex,naics,avg(work_value) AS WV from working_hour_data
where  sex = 'Females'
and geo <> 'Canada'
and year(reference_date)=2020
group by sex,naics) A group by naics,sex) whf
where whm.naics=whf.naics;
```

naics	male_wh	female_wh
Accommodation and food services [72]	1103.3300	1147.7700
Agriculture [111-112, 1100, 1151-1152]	869.8300	272.4700
Business, building and other support ser	1312.5600	706.1800
Construction [23]	4089.3800	523.5700
Educational services [61]	1123.6800	2207.4600
Finance, insurance, real estate, rental	2008.6300	1948.5200
Forestry, fishing, mining, quarrying, oi	979.5200	196.9900
Health care and social assistance [62]	1527.6300	5307.3600
Information, culture and recreation [51,	1149.5200	749.7500
Manufacturing [31-33]	4321.6500	1499.7700
Other services (except public administra	1146.7200	954.6800
Professional, scientific and technical s	3074.7000	1989.7700
Public administration [91]	1730.4200	1474.3700
Transportation and warehousing [48-49]	2556.4300	628.4600
Utilities [22]	368.7800	126.4700
Wholesale and retail trade [41, 44-45]	4609.9100	3270.4300