In [2]:

```
from pyspark.sql.functions import *
from pyspark.sql.types import *
from pyspark.sql import *

"""
import pandas as pd
#from IPython.core.display import display, HTML
#display(HTML('<h1>Hello, world!</h1>'))
"""
```

```
"\nimport pandas as pd\n#from IPython.core.display import display, HTML\n#dis
play(HTML('<h1>Hello, world!</h1>'))\n"
```

In [378]:

```
spark.sql("show tables").show()
```

```
+--------+---------------+-----------+
|database|      tableName|isTemporary|
+--------+---------------+-----------+
|        | adultincomedata|       true|
|        |adultincomedata1|       true|
|        |   customerdata|       true|
|        |             d1|       true|
|        |         demodf|       true|
|        |        demodf1|       true|
|        |       demodf22|       true|
|        |       demoview|       true|
|        |            df1|       true|
|        |         income|       true|
|        |       mergeddf|       true|
+--------+---------------+-----------+
```

Reading Consumer Dataset

In [57]:

```
cmplDF = spark.read.csv("s3://projectproposal2019/tools/cust_complains_all_thr
ough_3_31.csv", sep="\t")
cmplDF.cache()
cmplDF.take(1)
print(cmplDF.show(2))
print(cmplDF.printSchema())
```

```
+----------+--------------------+----------------+--------------------+------
--------------+----+----+--------------------+---+-----+----+----+----+------
----+--------------------+----+----+-------+
|       _c0|                 _c1|             _c2|                 _c3|
_c4| _c5| _c6|                 _c7|_c8|  _c9|_c10|_c11|_c12|      _c13|
_c14|_c15|_c16|   _c17|
+----------+--------------------+----------------+--------------------+------
--------------+----+----+--------------------+---+-----+----+----+----+------
----+--------------------+----+----+-------+
|05/03/2019|Credit reporting,...|Credit reporting|Problem with a cr...|Their
investigati...|null|null|Adler Wallach & A...| TN|37343|null|null| Web|05/03/
2019|Closed with expla...| Yes| N/A|3231390|
|05/03/2019|            Mortgage|    FHA mortgage|Applying for a mo...|
null|null|null|  FLAGSTAR BANK, FSB| FL|33032|null|null| Web|05/03/2019|
In progress| Yes| N/A|3231005|
+----------+--------------------+----------------+--------------------+------
--------------+----+----+--------------------+---+-----+----+----+----+------
----+--------------------+----+----+-------+
only showing top 2 rows

None
root
 |-- _c0: string (nullable = true)
 |-- _c1: string (nullable = true)
 |-- _c2: string (nullable = true)
 |-- _c3: string (nullable = true)
 |-- _c4: string (nullable = true)
 |-- _c5: string (nullable = true)
 |-- _c6: string (nullable = true)
 |-- _c7: string (nullable = true)
 |-- _c8: string (nullable = true)
 |-- _c9: string (nullable = true)
 |-- _c10: string (nullable = true)
 |-- _c11: string (nullable = true)
 |-- _c12: string (nullable = true)
 |-- _c13: string (nullable = true)
 |-- _c14: string (nullable = true)
 |-- _c15: string (nullable = true)
 |-- _c16: string (nullable = true)
 |-- _c17: string (nullable = true)

None
```

In [58]:

```
cmplDF=cmplDF.withColumnRenamed("_c0","DateReceived")\
.withColumnRenamed("_c1","Product")\
.withColumnRenamed("_c2","SubProduct")\
.withColumnRenamed("_c3",'Issue')\
.withColumnRenamed("_c4","SubIssue")\
.withColumnRenamed("_c5","ConsumerComplaintNarrative")\
.withColumnRenamed("_c6","CompanyPublicResponse")\
.withColumnRenamed("_c7",'Company')\
.withColumnRenamed("_c8",'State')\
.withColumnRenamed("_c9",'ZIPcode')\
.withColumnRenamed("_c10",'Tags')\
.withColumnRenamed("_c11",'ConsumerConsentProvided')\
.withColumnRenamed("_c12",'Submittedvia')\
.withColumnRenamed("_c13",'DateSentToCompany')\
.withColumnRenamed("_c14",'CompanyResponsetoConsumer')\
.withColumnRenamed("_c15",'TimelyResponse')\
.withColumnRenamed("_c16",'ConsumerDisputed')\
.withColumnRenamed("_c17",'ComplaintID')
```

In [59]:
```
cmplDF.columns
```

```
['DateReceived', 'Product', 'SubProduct', 'Issue', 'SubIssue', 'ConsumerCompl
aintNarrative', 'CompanyPublicResponse', 'Company', 'State', 'ZIPcode', 'Tag
s', 'ConsumerConsentProvided', 'Submittedvia', 'DateSentToCompany', 'CompanyR
esponsetoConsumer', 'TimelyResponse', 'ConsumerDisputed', 'ComplaintID']
```

In [60]:
```
cmplDF.createOrReplaceTempView("customerdata")
```

In [61]:
```
spark.sql("select distinct ComplaintID from customerdata group by ComplaintID,
ZIPcode").show(2)
```

```
+-----------+
|ComplaintID|
+-----------+
|    3216175|
|    3207829|
+-----------+
only showing top 2 rows
```

In [62]:
```
spark.sql("desc customerdata").show()
```

```
+--------------------+---------+-------+
|            col_name|data_type|comment|
+--------------------+---------+-------+
|        DateReceived|   string|   null|
|             Product|   string|   null|
|          SubProduct|   string|   null|
|               Issue|   string|   null|
|            SubIssue|   string|   null|
|ConsumerComplaint...|   string|   null|
|CompanyPublicResp...|   string|   null|
|             Company|   string|   null|
|               State|   string|   null|
|             ZIPcode|   string|   null|
|                Tags|   string|   null|
|ConsumerConsentPr...|   string|   null|
|        Submittedvia|   string|   null|
|   DateSentToCompany|   string|   null|
|CompanyResponseto...|   string|   null|
|      TimelyResponse|   string|   null|
|    ConsumerDisputed|   string|   null|
|         ComplaintID|   string|   null|
+--------------------+---------+-------+
```

In [63]:
```
customerdf=spark.sql("select * from customerdata")
```

In [65]:
```
customerdf.where(col("SubIssue").isNull()).count()
```

```
530524
```

In [66]:
```
fill_cols_vals = {"Company":"N/A","SubIssue": "N/A", "State" : "N/A", "SubProd
uct":"N/A", \
                  "ConsumerConsentProvided":"N/A", "Tags":"N/A", \
                 "ConsumerComplaintNarrative":"N/A" , "CompanyPublicResponse":
"N/A" , \
                  "ComplaintID" : "0"}
customerdf = customerdf.na.fill(fill_cols_vals)
```

In [67]:
```
customerdf.where(col("ConsumerConsentProvided")=="N/A").count()
```

```
589812
```

In [68]:
```
customerdf.columns
```

```
['DateReceived', 'Product', 'SubProduct', 'Issue', 'SubIssue', 'ConsumerCompl
aintNarrative', 'CompanyPublicResponse', 'Company', 'State', 'ZIPcode', 'Tag
s', 'ConsumerConsentProvided', 'Submittedvia', 'DateSentToCompany', 'CompanyR
esponsetoConsumer', 'TimelyResponse', 'ConsumerDisputed', 'ComplaintID']
```

In [69]:

```
customerdf.createOrReplaceTempView("customerdata")
```

In [70]:

```
spark.sql("select distinct Tags from customerdata ").show()
```

```
+--------------------+
|                Tags|
+--------------------+
|      Older American|
|                 N/A|
|       Servicemember|
|Older American, S...|
+--------------------+
```

In [199]:

```
spark.sql("select distinct Company, count(Product) as count from customerdata
 group by  \
Company order by count desc").show(10)
```

```
+--------------------+------+
|             Company| count|
+--------------------+------+
|       EQUIFAX, INC.|114721|
|Experian Informat...|103061|
|TRANSUNION INTERM...| 95807|
|BANK OF AMERICA, ...| 81910|
|WELLS FARGO & COM...| 70750|
|JPMORGAN CHASE & CO.| 60030|
|      CITIBANK, N.A.| 48925|
|CAPITAL ONE FINAN...| 34463|
|Navient Solutions...| 29227|
|OCWEN LOAN SERVIC...| 27731|
+--------------------+------+
only showing top 10 rows
```

In [198]:

```
spark.sql("select distinct Company, product, count(Product) as count from cust
omerdata group by Product, \
Company order by count desc").show(20)
```

```
+--------------------+--------------------+-----+
|             Company|             product|count|
+--------------------+--------------------+-----+
|       EQUIFAX, INC.|Credit reporting,...|64629|
|Experian Informat...|Credit reporting,...|55741|
|TRANSUNION INTERM...|Credit reporting,...|54208|
|       EQUIFAX, INC.|    Credit reporting|48124|
|Experian Informat...|    Credit reporting|45376|
|BANK OF AMERICA, ...|            Mortgage|42905|
|TRANSUNION INTERM...|    Credit reporting|39811|
|WELLS FARGO & COM...|            Mortgage|36629|
|OCWEN LOAN SERVIC...|            Mortgage|26503|
|Navient Solutions...|        Student loan|25107|
|JPMORGAN CHASE & CO.|            Mortgage|20985|
| NATIONSTAR MORTGAGE|            Mortgage|19609|
|      CITIBANK, N.A.|         Credit card|16817|
|BANK OF AMERICA, ...|Bank account or s...|13916|
|WELLS FARGO & COM...|Bank account or s...|13333|
|CAPITAL ONE FINAN...|         Credit card|12920|
|Ditech Financial LLC|            Mortgage|12894|
|ENCORE CAPITAL GR...|     Debt collection|10487|
|JPMORGAN CHASE & CO.|         Credit card|10373|
|JPMORGAN CHASE & CO.|Bank account or s...| 9816|
+--------------------+--------------------+-----+
only showing top 20 rows
```

```
In [72]: spark.sql("select distinct TimelyResponse, count(TimelyResponse) from customer
         data group by TimelyResponse").show()
```

```
+--------------+---------------------+
|TimelyResponse|count(TimelyResponse)|
+--------------+---------------------+
|          null|                    0|
|            No|                32088|
|           Yes|              1244602|
+--------------+---------------------+
```

```
In [73]: spark.sql("select distinct Company, ZIPcode, count(case when TimelyResponse='Y
         es' then 1 end) as resolved, \
         count(case when TimelyResponse='No' then 1 end) as Notresolved \
         from customerdata where ZIPcode!='' group by Company, ZIPcode order by resolve
         d desc").show()
```

```
+--------------------+-------+--------+-----------+
|             Company|ZIPcode|resolved|Notresolved|
+--------------------+-------+--------+-----------+
|       EQUIFAX, INC.|  300XX|     947|         24|
|       EQUIFAX, INC.|  330XX|     859|          9|
|Experian Informat...|  330XX|     805|          0|
|       EQUIFAX, INC.|  770XX|     779|         15|
|TRANSUNION INTERM...|  330XX|     759|          0|
|TRANSUNION INTERM...|  300XX|     703|          1|
|Experian Informat...|  300XX|     703|          0|
|       EQUIFAX, INC.|  331XX|     689|          8|
|Experian Informat...|  770XX|     669|          0|
|       EQUIFAX, INC.|  303XX|     662|         10|
|TRANSUNION INTERM...|  770XX|     648|          2|
|Experian Informat...|  331XX|     608|          0|
|TRANSUNION INTERM...|  331XX|     602|          3|
|       EQUIFAX, INC.|  606XX|     595|          8|
|Experian Informat...|  334XX|     542|          0|
|       EQUIFAX, INC.|  334XX|     531|         11|
|       EQUIFAX, INC.|  302XX|     523|          7|
|Experian Informat...|  303XX|     510|          0|
|TRANSUNION INTERM...|  303XX|     497|          0|
|TRANSUNION INTERM...|  606XX|     493|          0|
+--------------------+-------+--------+-----------+
only showing top 20 rows
```

Reading Adult Income

```
In [20]: adultIncome = spark.read.csv("s3://projectproposal2019/tools/adultincome.csv")
```

In [21]:

```
adultIncome.show(1)
adultIncome.printSchema()
```

```
+----------+---+---------+------+---------+---------------+--------------+---
-------+------------+----+------+------------+------------+--------------+---
-----------+------+-----------+
|       _c0|_c1|      _c2|   _c3|      _c4|            _c5|           _c6|
_c7|         _c8| _c9|  _c10|        _c11|        _c12|          _c13|
_c14|  _c15|       _c16|
+----------+---+---------+------+---------+---------------+--------------+---
-------+------------+----+------+------------+------------+--------------+---
-----------+------+-----------+
|customerID|age|workclass|fnlwgt|education|educational-num|marital-status|occ
upation|relationship|race|gender|capital-gain|capital-loss|hours-per-week|nat
ive-country|income|ComplaintID|
+----------+---+---------+------+---------+---------------+--------------+---
-------+------------+----+------+------------+------------+--------------+---
-----------+------+-----------+
only showing top 1 row

root
 |-- _c0: string (nullable = true)
 |-- _c1: string (nullable = true)
 |-- _c2: string (nullable = true)
 |-- _c3: string (nullable = true)
 |-- _c4: string (nullable = true)
 |-- _c5: string (nullable = true)
 |-- _c6: string (nullable = true)
 |-- _c7: string (nullable = true)
 |-- _c8: string (nullable = true)
 |-- _c9: string (nullable = true)
 |-- _c10: string (nullable = true)
 |-- _c11: string (nullable = true)
 |-- _c12: string (nullable = true)
 |-- _c13: string (nullable = true)
 |-- _c14: string (nullable = true)
 |-- _c15: string (nullable = true)
 |-- _c16: string (nullable = true)
```

In [22]:

```
adultIncome.createOrReplaceTempView("adultincomeData")
```

In [23]:

```
newincome= spark.sql("select * from adultincomeData \
where _c0 <>'customerID'")
```

In [24]:

```
newincome.show(2)
```

```
+----+---+-------+------+-------+---+------------------+-----------------+---
------+-----+----+----+----+----+-------------+-----+-------+
| _c0|_c1|    _c2|   _c3|    _c4|_c5|               _c6|              _c7|
_c8|  _c9|_c10|_c11|_c12|_c13|         _c14| _c15|   _c16|
+----+---+-------+------+-------+---+------------------+-----------------+---
------+-----+----+----+----+----+-------------+-----+-------+
|2200| 25|Private|226802|   11th|  7|     Never-married|Machine-op-inspct|Own
-child|Black|Male|   0|   0|  40|United-States|<=50K|3216175|
|2201| 38|Private| 89814|HS-grad|  9|Married-civ-spouse|  Farming-fishing|  H
usband|White|Male|   0|   0|  50|United-States|<=50K|3207829|
+----+---+-------+------+-------+---+------------------+-----------------+---
------+-----+----+----+----+----+-------------+-----+-------+
only showing top 2 rows
```

In [25]:
```
newincome=newincome.withColumnRenamed("_c0","customerID")\
.withColumnRenamed("_c1","age")\
.withColumnRenamed("_c2","workclass")\
.withColumnRenamed("_c3","fnlwgt")\
.withColumnRenamed("_c4","education")\
.withColumnRenamed("_c5","educational_num")\
.withColumnRenamed("_c6","marital_status")\
.withColumnRenamed("_c7","occupation")\
.withColumnRenamed("_c8","relationship")\
.withColumnRenamed("_c9","race")\
.withColumnRenamed("_c10","gender")\
.withColumnRenamed("_c11","capital_gain")\
.withColumnRenamed("_c12","capital_loss")\
.withColumnRenamed("_c13","hours_per_week")\
.withColumnRenamed("_c14","native_country")\
.withColumnRenamed("_c15","income")\
.withColumnRenamed("_c16","ComplaintID")
```

In [26]:
```
newincome.columns
```

```
['customerID', 'age', 'workclass', 'fnlwgt', 'education', 'educational_num',
'marital_status', 'occupation', 'relationship', 'race', 'gender', 'capital_ga
in', 'capital_loss', 'hours_per_week', 'native_country', 'income', 'Complaint
ID']
```

In [27]:
```
newincome.createOrReplaceTempView("adultincomeData1")
```

drop unwanted columns from dataframe

In [29]:
```
newincomedf= spark.sql("select customerID,age, workclass, \
education,marital_status,occupation,relationship, race, gender,income,Complain
tID from adultincomeData1")
```

In [30]:
```
newincomedf.show(2)
```

```
+----------+---+---------+---------+------------------+-----------------+----
--------+-----+------+------+-----------+
|customerID|age|workclass|education|    marital_status|       occupation|rela
tionship| race|gender|income|ComplaintID|
+----------+---+---------+---------+------------------+-----------------+----
--------+-----+------+------+-----------+
|      2200| 25|  Private|     11th|     Never-married|Machine-op-inspct|   O
wn-child|Black|  Male| <=50K|    3216175|
|      2201| 38|  Private|  HS-grad|Married-civ-spouse|  Farming-fishing|
Husband|White|  Male| <=50K|    3207829|
+----------+---+---------+---------+------------------+-----------------+----
--------+-----+------+------+-----------+
only showing top 2 rows
```

In [31]:
```
newincomedf.createOrReplaceTempView("income")
```

In [33]:
```
spark.sql("desc income").show()
```

```
+--------------+---------+-------+
|      col_name|data_type|comment|
+--------------+---------+-------+
|    customerID|   string|   null|
|           age|   string|   null|
|     workclass|   string|   null|
|     education|   string|   null|
|marital_status|   string|   null|
|    occupation|   string|   null|
|  relationship|   string|   null|
|          race|   string|   null|
|        gender|   string|   null|
|        income|   string|   null|
|   ComplaintID|   string|   null|
+--------------+---------+-------+
```

In [34]:

```
incomedf=spark.sql("select * from income")
```

In [35]:

```
incomedf.show(5)
```

```
+----------+---+---------+------------+------------------+-----------------+-
-----------+-----+------+------+-----------+
|customerID|age|workclass|   education|    marital_status|       occupation|r
elationship| race|gender|income|ComplaintID|
+----------+---+---------+------------+------------------+-----------------+-
-----------+-----+------+------+-----------+
|      2200| 25|  Private|        11th|     Never-married|Machine-op-inspct|
Own-child|Black|  Male| <=50K|    3216175|
|      2201| 38|  Private|     HS-grad|Married-civ-spouse|  Farming-fishing|
Husband|White|  Male| <=50K|    3207829|
|      2202| 28|Local-gov|  Assoc-acdm|Married-civ-spouse|  Protective-serv|
Husband|White|  Male|  >50K|    3205615|
|      2203| 44|  Private|Some-college|Married-civ-spouse|Machine-op-inspct|
Husband|Black|  Male|  >50K|    3201821|
|      2204| 18|        ?|Some-college|     Never-married|                ?|
Own-child|White|Female| <=50K|    3200629|
+----------+---+---------+------------+------------------+-----------------+-
-----------+-----+------+------+-----------+
only showing top 5 rows
```

In [36]:

```
spark.sql("select sum(age) from income").show(5)
```

```
+------------------------+
|sum(CAST(age AS DOUBLE))|
+------------------------+
|               1159305.0|
+------------------------+
```

find correlation between age, income and occupation of a person

In [37]:

```
from pyspark.sql.functions import countDistinct, approx_count_distinct, count,
sum, mean, round
incomedf.groupBy(col("occupation"), col("income")).agg(round(mean("age"),3).al
ias("Average-age"))\
.sort("Average-age", ascending=False).show(10)
```

```
+-----------------+------+-----------+
|       occupation|income|Average-age|
+-----------------+------+-----------+
|                ?|  >50K|       56.0|
|  Farming-fishing|  >50K|     46.471|
|  Priv-house-serv| <=50K|     45.107|
| Transport-moving|  >50K|     45.097|
|  Exec-managerial|  >50K|     44.713|
|     Adm-clerical|  >50K|     44.224|
|            Sales|  >50K|     44.134|
|   Prof-specialty|  >50K|     43.761|
|Handlers-cleaners|  >50K|     43.528|
|Machine-op-inspct|  >50K|     43.414|
+-----------------+------+-----------+
only showing top 10 rows
```

In [38]:
```
incomedf.groupBy(col("occupation"), col("income")) \
.agg(round(mean("age"),3).alias("Average-age"))\
.sort("Average-age", ascending=True).show(10)
```

```
+-----------------+------+-----------+
|       occupation|income|Average-age|
+-----------------+------+-----------+
|     Armed-Forces| <=50K|       25.0|
|Handlers-cleaners| <=50K|     32.054|
|     Tech-support| <=50K|     34.832|
|    Other-service| <=50K|     34.839|
|            Sales| <=50K|     35.002|
|     Adm-clerical| <=50K|     36.442|
|Machine-op-inspct| <=50K|     37.054|
|   Prof-specialty| <=50K|     37.706|
|                ?| <=50K|     37.713|
|     Craft-repair| <=50K|     37.923|
+-----------------+------+-----------+
only showing top 10 rows
```

In [39]:
```
spark.sql("SELECT gender, count(ComplaintID) as totalcomplaintRegistered \
    FROM income GROUP BY gender").show()
```

```
+------+------------------------+
|gender|totalcomplaintRegistered|
+------+------------------------+
|Female|                    9944|
|  Male|                   20055|
+------+------------------------+
```

merge customerdata and income

In [40]:
```
spark.sql("desc income").show()
```

```
+--------------+---------+-------+
|      col_name|data_type|comment|
+--------------+---------+-------+
|    customerID|   string|   null|
|           age|   string|   null|
|     workclass|   string|   null|
|     education|   string|   null|
|marital_status|   string|   null|
|    occupation|   string|   null|
|  relationship|   string|   null|
|          race|   string|   null|
|        gender|   string|   null|
|        income|   string|   null|
|   ComplaintID|   string|   null|
+--------------+---------+-------+
```

In [81]:
```
mrgeddf = spark.sql("select c.DateReceived, c.Product , c.SubProduct,  c.Issu
e, c.SubIssue, c.ConsumerComplaintNarrative, c.CompanyPublicResponse \
, c.Company, c.State, c.ZIPcode, c.Tags, c.ConsumerConsentProvided,  c.Submitt
edvia, \
c.DateSentToCompany, c.CompanyResponsetoConsumer, c.TimelyResponse, c.Consumer
Disputed, c.ComplaintID, \
i.customerID, i.age, i.workclass, i.education,  i.marital_status, i.occupatio
n, \
i.relationship, i.race, i.gender,  i.income \
from income i inner join customerdata c on i.ComplaintID = c.ComplaintID ")
```

In [82]:
```
mrgeddf.count()
```

```
879914
```

In [83]:
```
mrgeddf.createOrReplaceTempView("mergedDF")
```

In [195]:

```
spark.sql("select distinct round((count(case when TimelyResponse='Yes' then 1
 end)/count(*))*100,3)as Percentageresolved, \
round((count(case when TimelyResponse='No' then 1 end)/count(*))*100,3) as Per
centageNotresolved from customerdata").show(3)
```

```
+------------------+---------------------+
|Percentageresolved|PercentageNotresolved|
+------------------+---------------------+
|             97.48|                2.513|
+------------------+---------------------+
```

In [85]:

```
spark.sql("select * from mergedDF").show(2)
```

```
+------------+---------------+----------------+--------------------+---------
-----------+--------------------------+---------------------+----------------
----+-----+-------+----+-----------------------+------------+----------------
-+-------------------------+--------------+----------------+-----------+-----
-----+---+---------+---------+------------------+-----------------+----------
--+-----+------+------+
|DateReceived|        Product|      SubProduct|               Issue|
SubIssue|ConsumerComplaintNarrative|CompanyPublicResponse|             Compan
y|State|ZIPcode|Tags|ConsumerConsentProvided|Submittedvia|DateSentToCompany|C
ompanyResponsetoConsumer|TimelyResponse|ConsumerDisputed|ComplaintID|customer
ID|age|workclass|education|    marital_status|       occupation|relationship|
race|gender|income|
+------------+---------------+----------------+--------------------+---------
-----------+--------------------------+---------------------+----------------
----+-----+-------+----+-----------------------+------------+----------------
-+-------------------------+--------------+----------------+-----------+-----
-----+---+---------+---------+------------------+-----------------+----------
--+-----+------+------+
|  05/03/2019|Debt collection|Credit card debt|Attempts to colle...|Debt was
already ...|                       N/A|                  N/A|NAVY FEDERAL CRE
D...|   AL|  358XX| N/A|                    N/A|         Web|       05/03/201
9|              In progress|           Yes|             N/A|    3231618|
16238| 35|  Private|Assoc-voc|Married-civ-spouse|Machine-op-inspct|     Husba
nd|White|  Male| <=50K|
|  05/02/2019|Debt collection|   I do not know|Threatened to con...|Talked to
a third...|                       N/A|                  N/A|Cedars Business
S...|   VA|  20171| N/A|                    N/A|         Web|       05/02/201
9|     Closed with expla...|           Yes|             N/A|    3230073|
16758| 29|  Private|  HS-grad|          Divorced|     Craft-repair|   Own-chi
ld|White|  Male| <=50K|
+------------+---------------+----------------+--------------------+---------
-----------+--------------------------+---------------------+----------------
----+-----+-------+----+-----------------------+------------+----------------
-+-------------------------+--------------+----------------+-----------+-----
-----+---+---------+---------+------------------+-----------------+----------
--+-----+------+------+
only showing top 2 rows
```

In [86]:

```
spark.sql("desc mergedDF").show(45)
```

```
+--------------------+---------+-------+
|            col_name|data_type|comment|
+--------------------+---------+-------+
|        DateReceived|   string|   null|
|             Product|   string|   null|
|          SubProduct|   string|   null|
|               Issue|   string|   null|
|            SubIssue|   string|   null|
|ConsumerComplaint...|   string|   null|
|CompanyPublicResp...|   string|   null|
|             Company|   string|   null|
|               State|   string|   null|
|             ZIPcode|   string|   null|
|                Tags|   string|   null|
|ConsumerConsentPr...|   string|   null|
|         Submittedvia|   string|   null|
|   DateSentToCompany|   string|   null|
|CompanyResponseto...|   string|   null|
|      TimelyResponse|   string|   null|
|    ConsumerDisputed|   string|   null|
|         ComplaintID|   string|   null|
|          customerID|   string|   null|
|                 age|   string|   null|
|           workclass|   string|   null|
|           education|   string|   null|
|      marital_status|   string|   null|
|          occupation|   string|   null|
|        relationship|   string|   null|
|                race|   string|   null|
|              gender|   string|   null|
|              income|   string|   null|
+--------------------+---------+-------+
```

In [87]:

```
spark.sql("SELECT gender, count(ComplaintID) as totalcomplaintRegistered \
    FROM mergedDF GROUP BY gender").show()
```

```
+------+------------------------+
|gender|totalcomplaintRegistered|
+------+------------------------+
|Female|                  292229|
|  Male|                  587685|
+------+------------------------+
```

In [88]:

```
mrgeddf.groupBy("gender").count().show()
```

```
+------+------+
|gender| count|
+------+------+
|Female|292229|
|  Male|587685|
+------+------+
```

In [146]:

```
fill_cols_vals = {"Submittedvia":"N/A"}
mrgeddf = mrgeddf.na.fill(fill_cols_vals)
```

In [147]:

```
mrgeddf.select(col("Submittedvia"), col("occupation")) \
.sort("Submittedvia", ascending=False).show()
```

```
+------------+-----------------+
|Submittedvia|       occupation|
+------------+-----------------+
|         Web|     Adm-clerical|
|         Web|   Prof-specialty|
|         Web|Handlers-cleaners|
|         Web|     Adm-clerical|
|         Web|   Prof-specialty|
|         Web|     Adm-clerical|
|         Web|     Craft-repair|
|         Web|Handlers-cleaners|
|         Web|     Adm-clerical|
|         Web|  Exec-managerial|
|         Web|     Tech-support|
|         Web|     Adm-clerical|
|         Web|     Adm-clerical|
|         Web|                ?|
|         Web|     Craft-repair|
|         Web|   Prof-specialty|
|         Web|            Sales|
|         Web|  Protective-serv|
|         Web|            Sales|
|         Web|    Other-service|
+------------+-----------------+
only showing top 20 rows
```

In [149]:

```
df= mrgeddf.groupBy(col("Submittedvia"), col("income")).agg(count("Submittedvi
a").alias("total-Submittedvia"))\
.sort("total-Submittedvia", ascending=False).show()
```

```
+------------+------+------------------+
|Submittedvia|income|total-Submittedvia|
+------------+------+------------------+
|         N/A| <=50K|            656524|
|         N/A|  >50K|            203390|
|         Web| <=50K|             11171|
|         Web|  >50K|              3497|
|    Referral| <=50K|              2132|
|       Phone| <=50K|               879|
| Postal mail| <=50K|               810|
|    Referral|  >50K|               624|
|       Phone|  >50K|               311|
| Postal mail|  >50K|               248|
|         Fax| <=50K|               240|
|         Fax|  >50K|                80|
|       Email|  >50K|                 5|
|       Email| <=50K|                 3|
+------------+------+------------------+
```

In [150]:
```
spark.sql("select Submittedvia,count(Submittedvia) as countSubmittedVia, \
income from mergedDF group by income, Submittedvia  order by countSubmittedVia
desc").show()
```

```
+------------+-----------------+------+
|Submittedvia|countSubmittedVia|income|
+------------+-----------------+------+
|         Web|            11171| <=50K|
|         Web|             3497|  >50K|
|    Referral|             2132| <=50K|
|       Phone|              879| <=50K|
| Postal mail|              810| <=50K|
|    Referral|              624|  >50K|
|       Phone|              311|  >50K|
| Postal mail|              248|  >50K|
|         Fax|              240| <=50K|
|         Fax|               80|  >50K|
|       Email|                5|  >50K|
|       Email|                3| <=50K|
|        null|                0| <=50K|
|        null|                0|  >50K|
+------------+-----------------+------+
```

In [185]:
```
df=spark.sql("select distinct income, count(*) from mergedDF group by income")
df.show()
```

```
+------+--------+
|income|count(1)|
+------+--------+
| <=50K|  671759|
|  >50K|  208155|
+------+--------+
```

Reading third dataset- demographics

In [202]:
```
spark.sql("select * from customerdata").show(2)
```

```
+------------+--------------------+----------------+--------------------+----
----------------+--------------------------+---------------------+----------
---------+-----+-------+----+-----------------------+------------+----------
------+-------------------------+--------------+----------------+-----------+
|DateReceived|             Product|      SubProduct|               Issue|
SubIssue|ConsumerComplaintNarrative|CompanyPublicResponse|             Compan
y|State|ZIPcode|Tags|ConsumerConsentProvided|Submittedvia|DateSentToCompany|C
ompanyResponsetoConsumer|TimelyResponse|ConsumerDisputed|ComplaintID|
+------------+--------------------+----------------+--------------------+----
----------------+--------------------------+---------------------+----------
---------+-----+-------+----+-----------------------+------------+----------
------+-------------------------+--------------+----------------+-----------+
|  05/03/2019|Credit reporting,...|Credit reporting|Problem with a cr...|Thei
r investigati...|                       N/A|                  N/A|Adler Walla
ch & A...|   TN|  37343| N/A|                    N/A|         Web|       05/0
3/2019|     Closed with expla...|           Yes|             N/A|    3231390|
|  05/03/2019|            Mortgage|    FHA mortgage|Applying for a mo...|
N/A|                       N/A|                  N/A|  FLAGSTAR BANK, FSB|
FL|  33032| N/A|                    N/A|         Web|       05/03/2019|
In progress|           Yes|             N/A|    3231005|
+------------+--------------------+----------------+--------------------+----
----------------+--------------------------+---------------------+----------
---------+-----+-------+----+-----------------------+------------+----------
------+-------------------------+--------------+----------------+-----------+
only showing top 2 rows
```

In [242]:

```
demodf1 = spark.read.csv("s3://projectproposal2019/tools/Demographics.csv")
demodf1.cache()
demodf1.take(1)
print(demodf1.show(2))
print(demodf1.printSchema())
```

```
+-----------+--------------------+-----------+--------------------+------+---
------------+--------+-----------------+-------------------+--------------+--
---------+-------+--------+--------------------+--------+----------+
|        _c0|                 _c1|        _c2|                 _c3|   _c4|
_c5|     _c6|              _c7|                _c8|           _c9|       _c10
|   _c11|    _c12|                _c13|    _c14|      _c15|
+-----------+--------------------+-----------+--------------------+------+---
------------+--------+-----------------+-------------------+--------------+--
---------+-------+--------+--------------------+--------+----------+
|Panelist ID|Combined Pre-Tax ...|Family Size|Type of Residenti...|COUNTY|Mal
eWorkingHour|AgeGroup|FemaleWorkingHour|Children Group Code|Marital Status|Nu
mberofTVs|ZIPCODE|FIPSCODE|IRI Geography Number|EXT_FACT|customerID|
|    1100016|                   3|          3|                   2|     C|
4|       6|                4|                  8|             2|          3|
1201|   25003|                   1|       1|      2200|
+-----------+--------------------+-----------+--------------------+------+---
------------+--------+-----------------+-------------------+--------------+--
---------+-------+--------+--------------------+--------+----------+
only showing top 2 rows

None
root
 |-- _c0: string (nullable = true)
 |-- _c1: string (nullable = true)
 |-- _c2: string (nullable = true)
 |-- _c3: string (nullable = true)
 |-- _c4: string (nullable = true)
 |-- _c5: string (nullable = true)
 |-- _c6: string (nullable = true)
 |-- _c7: string (nullable = true)
 |-- _c8: string (nullable = true)
 |-- _c9: string (nullable = true)
 |-- _c10: string (nullable = true)
 |-- _c11: string (nullable = true)
 |-- _c12: string (nullable = true)
 |-- _c13: string (nullable = true)
 |-- _c14: string (nullable = true)
 |-- _c15: string (nullable = true)

None
```

In [250]:

```
demodf1.createOrReplaceTempView("demodf")
```

In [279]:

```
demodf1.show(2)
```

```
+-----------+--------------------+-----------+--------------------+------+---
------------+--------+-----------------+-------------------+--------------+--
---------+-------+--------+--------------------+--------+----------+
|        _c0|                 _c1|        _c2|                 _c3|   _c4|
_c5|     _c6|              _c7|                _c8|           _c9|       _c10
|   _c11|    _c12|                _c13|    _c14|      _c15|
+-----------+--------------------+-----------+--------------------+------+---
------------+--------+-----------------+-------------------+--------------+--
---------+-------+--------+--------------------+--------+----------+
|Panelist ID|Combined Pre-Tax ...|Family Size|Type of Residenti...|COUNTY|Mal
eWorkingHour|AgeGroup|FemaleWorkingHour|Children Group Code|Marital Status|Nu
mberofTVs|ZIPCODE|FIPSCODE|IRI Geography Number|EXT_FACT|customerID|
|    1100016|                   3|          3|                   2|     C|
4|       6|                4|                  8|             2|          3|
1201|   25003|                   1|       1|      2200|
+-----------+--------------------+-----------+--------------------+------+---
------------+--------+-----------------+-------------------+--------------+--
---------+-------+--------+--------------------+--------+----------+
only showing top 2 rows
```

In [257]:
```
demodf=demodf1.withColumnRenamed("_c0","PanelistID") \
.withColumnRenamed("_c1","preTaxIncome") \
.withColumnRenamed("_c2","familysize") \
.withColumnRenamed("_c3","TypeofResidentialPossession") \
.withColumnRenamed("_c4","COUNTY") \
.withColumnRenamed("_c5","MaleWorkingHour") \
.withColumnRenamed("_c6","AgeGroup") \
.withColumnRenamed("_c7","FemaleWorkingHour") \
.withColumnRenamed("_c8","ChildrenGroup") \
.withColumnRenamed("_c9","MaritalStatus") \
.withColumnRenamed("_c10","NumberofTVs") \
.withColumnRenamed("_c11","ZIPCODE") \
.withColumnRenamed("_c12","FIPSCODE") \
.withColumnRenamed("_c13","IRIGeographyNumber") \
.withColumnRenamed("_c14","EXT_FACT") \
.withColumnRenamed("_c15","customerID")
```

In [259]:
```
demodf.columns
```

```
['PanelistID', 'preTaxIncome', 'familysize', 'TypeofResidentialPossession',
'COUNTY', 'MaleWorkingHour', 'AgeGroup', 'FemaleWorkingHour', 'ChildrenGrou
p', 'MaritalStatus', 'NumberofTVs', 'ZIPCODE', 'FIPSCODE', 'IRIGeographyNumbe
r', 'EXT_FACT', 'customerID']
```

In [260]:
```
demodf.createOrReplaceTempView("demodf1")
```

In [261]:
```
demodf.show(5)
```

```
+-----------+--------------------+-----------+---------------------------+---
---+---------------+--------+-----------------+-------------------+----------
----+-----------+-------+--------+--------------------+--------+----------+
| PanelistID|        preTaxIncome| familysize|TypeofResidentialPossession|COU
NTY|MaleWorkingHour|AgeGroup|FemaleWorkingHour|      ChildrenGroup| MaritalSt
atus|NumberofTVs|ZIPCODE|FIPSCODE|  IRIGeographyNumber|EXT_FACT|customerID|
+-----------+--------------------+-----------+---------------------------+---
---+---------------+--------+-----------------+-------------------+----------
----+-----------+-------+--------+--------------------+--------+----------+
|Panelist ID|Combined Pre-Tax ...|Family Size|       Type of Residenti...|COU
NTY|MaleWorkingHour|AgeGroup|FemaleWorkingHour|Children Group Code|Marital St
atus|NumberofTVs|ZIPCODE|FIPSCODE|IRI Geography Number|EXT_FACT|customerID|
|    1100016|                   3|          3|                          2|
C|              4|       6|                4|                  8|
2|          3|   1201|   25003|                   1|       1|      2200|
|    1100032|                   5|          2|                          2|
C|              7|       5|                3|                  3|
1|          1|   1201|   25003|                   1|       1|      2201|
|    1100057|                  10|          2|                          2|
C|              4|       6|                3|                  8|
2|          3|   1201|   25003|                   1|       1|      2202|
|    1100156|                   7|          2|                          2|
C|              4|       6|                4|                  8|
2|          1|   1201|   25003|                   1|       1|      2203|
+-----------+--------------------+-----------+---------------------------+---
---+---------------+--------+-----------------+-------------------+----------
----+-----------+-------+--------+--------------------+--------+----------+
only showing top 5 rows
```

In [273]:
```
demodf2 = spark.sql("select * from demodf1 \
where PanelistID <>'Panelist ID'")
```

In [274]:
```
demodf2.columns
```

```
['PanelistID', 'preTaxIncome', 'familysize', 'TypeofResidentialPossession',
'COUNTY', 'MaleWorkingHour', 'AgeGroup', 'FemaleWorkingHour', 'ChildrenGrou
p', 'MaritalStatus', 'NumberofTVs', 'ZIPCODE', 'FIPSCODE', 'IRIGeographyNumbe
r', 'EXT_FACT', 'customerID']
```

In [275]:
```
demodf2.createOrReplaceTempView("demodf22")
```

In [278]:

```
demodf2.show(2)
```

```
+----------+------------+----------+---------------------------+------+------
---------+--------+-----------------+-------------+-------------+-----------+
-------+--------+------------------+--------+----------+
|PanelistID|preTaxIncome|familysize|TypeofResidentialPossession|COUNTY|MaleWo
rkingHour|AgeGroup|FemaleWorkingHour|ChildrenGroup|MaritalStatus|NumberofTVs|
ZIPCODE|FIPSCODE|IRIGeographyNumber|EXT_FACT|customerID|
+----------+------------+----------+---------------------------+------+------
---------+--------+-----------------+-------------+-------------+-----------+
-------+--------+------------------+--------+----------+
|   1100016|           3|         3|                          2|     C|
4|       6|                4|            8|            2|          3|   1201|
25003|                 1|       1|      2200|
|   1100032|           5|         2|                          2|     C|
7|       5|                3|            3|            1|          1|   1201|
25003|                 1|       1|      2201|
+----------+------------+----------+---------------------------+------+------
---------+--------+-----------------+-------------+-------------+-----------+
-------+--------+------------------+--------+----------+
only showing top 2 rows
```

In [288]:

```
demodf2.agg(sum("MaleWorkingHour").alias("total-MaleWorkingHour") ,sum("Female
WorkingHour").alias("total-FemaleWorkingHour"))\
.sort("total-MaleWorkingHour", ascending=False).show()
```

```
+--------------------+----------------------+
|total-MaleWorkingHour|total-FemaleWorkingHour|
+--------------------+----------------------+
|             28374.0|               22289.0|
+--------------------+----------------------+
```

In [289]:

```
demodf2.groupBy("familysize").agg(sum("MaleWorkingHour").alias("total-MaleWork
ingHour") \
                                  ,sum("FemaleWorkingHour").alias("total-Femal
eWorkingHour"))\
.sort("total-MaleWorkingHour", ascending=False).show()
```

```
+----------+--------------------+----------------------+
|familysize|total-MaleWorkingHour|total-FemaleWorkingHour|
+----------+--------------------+----------------------+
|         1|              9883.0|                6236.0|
|         2|              9793.0|                8534.0|
|         3|              3851.0|                3009.0|
|         4|              3043.0|                2837.0|
|         5|              1288.0|                1170.0|
|         6|               516.0|                 503.0|
+----------+--------------------+----------------------+
```

In [290]:

```
spark.sql("select * from demodf22").show(2)
```

```
+----------+------------+----------+---------------------------+------+------
---------+--------+-----------------+-------------+-------------+-----------+
-------+--------+------------------+--------+----------+
|PanelistID|preTaxIncome|familysize|TypeofResidentialPossession|COUNTY|MaleWo
rkingHour|AgeGroup|FemaleWorkingHour|ChildrenGroup|MaritalStatus|NumberofTVs|
ZIPCODE|FIPSCODE|IRIGeographyNumber|EXT_FACT|customerID|
+----------+------------+----------+---------------------------+------+------
---------+--------+-----------------+-------------+-------------+-----------+
-------+--------+------------------+--------+----------+
|   1100016|           3|         3|                          2|     C|
4|       6|                4|            8|            2|          3|   1201|
25003|                 1|       1|      2200|
|   1100032|           5|         2|                          2|     C|
7|       5|                3|            3|            1|          1|   1201|
25003|                 1|       1|      2201|
+----------+------------+----------+---------------------------+------+------
---------+--------+-----------------+-------------+-------------+-----------+
-------+--------+------------------+--------+----------+
only showing top 2 rows
```

Merge all 3 datasets

In [353]:

```
allmergeddf = spark.sql("select c.DateReceived, c.Product , c.SubProduct,  c.Issue, c.SubIssue, \
c.ConsumerComplaintNarrative, c.CompanyPublicResponse \
, c.Company, c.State, c.ZIPcode, c.Tags, c.ConsumerConsentProvided,  c.Submittedvia, \
c.DateSentToCompany, c.CompanyResponsetoConsumer, c.TimelyResponse, c.ConsumerDisputed, c.ComplaintID, \
i.customerID, i.age, i.workclass, i.education,  i.marital_status, i.occupation, \
i.relationship, i.race, i.gender,  i.income, \
d.preTaxIncome, d.familysize, d.TypeofResidentialPossession, d.COUNTY, d.MaleWorkingHour, d.AgeGroup, d.FemaleWorkingHour, \
d.ChildrenGroup, d.NumberofTVs, d.FIPSCODE, d.IRIGeographyNumber, d.EXT_FACT \
from income i, customerdata c , demodf22 d \
where i.ComplaintID = c.ComplaintID \
and i.customerID = d.customerID")
```

In [298]:

```
allmergeddf.show(2)
```

```
+------------+--------------------+----------------+--------------------+----
----------------+--------------------------+---------------------+-----------
---------+-----+-------+----+-----------------------+------------+-----------
------+-------------------------+--------------+----------------+-----------+
----------+---+---------+---------+------------------+--------------+--------
-----+-----+------+------+------------+----------+---------------------------
+------+---------------+--------+-----------------+-------------+-----------+
-------+--------+------------------+--------+
|DateReceived|             Product|      SubProduct|               Issue|
SubIssue|ConsumerComplaintNarrative|CompanyPublicResponse|             Compan
y|State|ZIPcode|Tags|ConsumerConsentProvided|Submittedvia|DateSentToCompany|C
ompanyResponsetoConsumer|TimelyResponse|ConsumerDisputed|ComplaintID|customer
ID|age|workclass|education|    marital_status|    occupation| relationship| r
ace|gender|income|preTaxIncome|familysize|TypeofResidentialPossession|COUNTY|
MaleWorkingHour|AgeGroup|FemaleWorkingHour|ChildrenGroup|NumberofTVs|ZIPCODE|
FIPSCODE|IRIGeographyNumber|EXT_FACT|
+------------+--------------------+----------------+--------------------+----
----------------+--------------------------+---------------------+-----------
---------+-----+-------+----+-----------------------+------------+-----------
------+-------------------------+--------------+----------------+-----------+
----------+---+---------+---------+------------------+--------------+--------
-----+-----+------+------+------------+----------+---------------------------
+------+---------------+--------+-----------------+-------------+-----------+
-------+--------+------------------+--------+
|  04/26/2019|     Debt collection|   I do not know|Communication tac...|Freq
uent or repea...|                       N/A|                  N/A|Richland Bu
reau o...|   TN|   null| N/A|                    N/A|         Web|       04/2
6/2019|     Closed with expla...|           Yes|             N/A|    3223828|
8595| 35|  Private|Bachelors|Married-civ-spouse|  Adm-clerical|      Husband|
White|  Male| <=50K|           9|         3|                          2|
D|              4|       4|                3|            3|          3|  5470
1|   55033|                 3|       1|
|  04/26/2019|Credit reporting,...|Credit reporting|Incorrect informa...|Info
rmation belon...|                       N/A| Company has respo...|Experian In
format...|   FL|  32209| N/A|   Consent not provided|         Web|       04/2
6/2019|     Closed with non-m...|           Yes|             N/A|    3223564|
7233| 35|  Private|  HS-grad|          Divorced|Prof-specialty|Not-in-family|
White|  Male| <=50K|           9|         2|                          1|
D|              4|       6|                1|            8|          3|  5470
1|   55033|                 3|       1|
+------------+--------------------+----------------+--------------------+----
----------------+--------------------------+---------------------+-----------
---------+-----+-------+----+-----------------------+------------+-----------
------+-------------------------+--------------+----------------+-----------+
----------+---+---------+---------+------------------+--------------+--------
-----+-----+------+------+------------+----------+---------------------------
+------+---------------+--------+-----------------+-------------+-----------+
-------+--------+------------------+--------+
only showing top 2 rows
```

In [306]:

```
allmergeddf.select(col("Product"), col("AgeGroup")).sort(col("Product"), ascen
ding=True).show()
```

```
+--------------------+--------+
|             Product|AgeGroup|
+--------------------+--------+
|Bank account or s...|       7|
|Bank account or s...|       4|
|Bank account or s...|       4|
|Bank account or s...|       5|
|Bank account or s...|       5|
|Bank account or s...|       4|
|Bank account or s...|       6|
|Bank account or s...|       5|
|Bank account or s...|       4|
|Bank account or s...|       3|
|Bank account or s...|       6|
|Bank account or s...|       6|
|Bank account or s...|       4|
|Bank account or s...|       7|
|Bank account or s...|       5|
|Bank account or s...|       3|
|Bank account or s...|       5|
|Bank account or s...|       5|
|Bank account or s...|       6|
|Bank account or s...|       2|
+--------------------+--------+
only showing top 20 rows
```

In [324]:

```
allmergeddf.groupBy(col("Product"), col("AgeGroup")) \
.agg(count(col("Product")).alias("total-Product"))  \
.sort(col("total-Product"), ascending=False).show()
```

```
+--------------------+--------+-------------+
|             Product|AgeGroup|total-Product|
+--------------------+--------+-------------+
|            Mortgage|       6|          425|
|     Debt collection|       6|          346|
|Credit reporting,...|       6|          337|
|            Mortgage|       4|          296|
|Credit reporting,...|       4|          278|
|     Debt collection|       4|          277|
|            Mortgage|       5|          257|
|            Mortgage|       3|          249|
|     Debt collection|       5|          240|
|Credit reporting,...|       5|          213|
|    Credit reporting|       6|          201|
|     Debt collection|       3|          197|
|Credit reporting,...|       3|          182|
|    Credit reporting|       4|          146|
|    Credit reporting|       5|          139|
|Bank account or s...|       6|          131|
|    Credit reporting|       3|          118|
|         Credit card|       6|          117|
|Bank account or s...|       4|          114|
|            Mortgage|       7|          110|
+--------------------+--------+-------------+
only showing top 20 rows
```

In [351]:

```
allmergeddf.groupBy(col("Product"), col("AgeGroup")) \
.agg(count(col("Product")).alias("total-Product"))  \
.where(col("total-Product")>250).where(col("AgeGroup")>=5)  \
.orderBy(col("AgeGroup"), ascending=False) \
.show()
```

```
+--------------------+--------+-------------+
|             Product|AgeGroup|total-Product|
+--------------------+--------+-------------+
|            Mortgage|       6|          425|
|Credit reporting,...|       6|          337|
|     Debt collection|       6|          346|
|            Mortgage|       5|          257|
+--------------------+--------+-------------+
```

In [352]:

```
allmergeddf.groupBy(col("Product"), col("AgeGroup")) \
.agg(count(col("Product")).alias("total-Product"))  \
.where(col("total-Product")>200) .where(col("AgeGroup")<5)\
.orderBy(col("AgeGroup"), ascending=True) \
.show()
```

```
+--------------------+--------+-------------+
|             Product|AgeGroup|total-Product|
+--------------------+--------+-------------+
|            Mortgage|       3|          249|
|            Mortgage|       4|          296|
|     Debt collection|       4|          277|
|Credit reporting,...|       4|          278|
+--------------------+--------+-------------+
```

In [369]:

```
allmergeddf.groupBy(col("income"), col("familysize")) \
.agg(sum(col("MaleWorkingHour")).alias("HoursMalework"), \
    sum(col("FemaleWorkingHour")).alias("HoursFemalework"), \
    count("income").alias("total"))\
.sort(col("HoursFemalework"), col("HoursMalework"), ascending= False).show()
```

```
+------+----------+-------------+---------------+-----+
|income|familysize|HoursMalework|HoursFemalework|total|
+------+----------+-------------+---------------+-----+
| <=50K|         2|       7581.0|         6597.0| 1990|
| <=50K|         1|       7410.0|         4674.0| 1126|
| <=50K|         3|       2995.0|         2285.0|  745|
| <=50K|         4|       2335.0|         2130.0|  737|
|  >50K|         2|       2212.0|         1937.0|  577|
|  >50K|         1|       2473.0|         1562.0|  377|
| <=50K|         5|        953.0|          858.0|  294|
|  >50K|         3|        856.0|          724.0|  234|
|  >50K|         4|        708.0|          707.0|  223|
| <=50K|         6|        386.0|          379.0|  122|
|  >50K|         5|        335.0|          312.0|  101|
|  >50K|         6|        130.0|          124.0|   39|
+------+----------+-------------+---------------+-----+
```

In [377]:

```
allmergeddf.groupBy(col("Issue"),col("occupation")) \
.agg(count("Issue").alias("total")) \
.where(col("Issue")!= "Incorrect information")\
.sort(col("total"), ascending=False).show()
```

```
+--------------------+---------------+-----+
|               Issue|     occupation|total|
+--------------------+---------------+-----+
|Incorrect informa...| Prof-specialty|   96|
|Loan modification...|   Adm-clerical|   91|
|Incorrect informa...|Exec-managerial|   90|
|Incorrect informa...| Prof-specialty|   83|
|Loan modification...|   Craft-repair|   83|
|Incorrect informa...|          Sales|   82|
|Incorrect informa...|   Craft-repair|   80|
|Incorrect informa...|  Other-service|   73|
|Incorrect informa...|   Craft-repair|   72|
|Loan modification...|  Other-service|   71|
|Loan modification...|Exec-managerial|   69|
|Incorrect informa...|   Adm-clerical|   63|
|Loan modification...| Prof-specialty|   60|
|Loan modification...|          Sales|   60|
|Incorrect informa...|   Adm-clerical|   58|
|Loan servicing, p...|Exec-managerial|   57|
|Incorrect informa...|          Sales|   54|
|Incorrect informa...|Exec-managerial|   54|
|Loan servicing, p...|   Adm-clerical|   54|
|Incorrect informa...|              ?|   52|
+--------------------+---------------+-----+
only showing top 20 rows
```