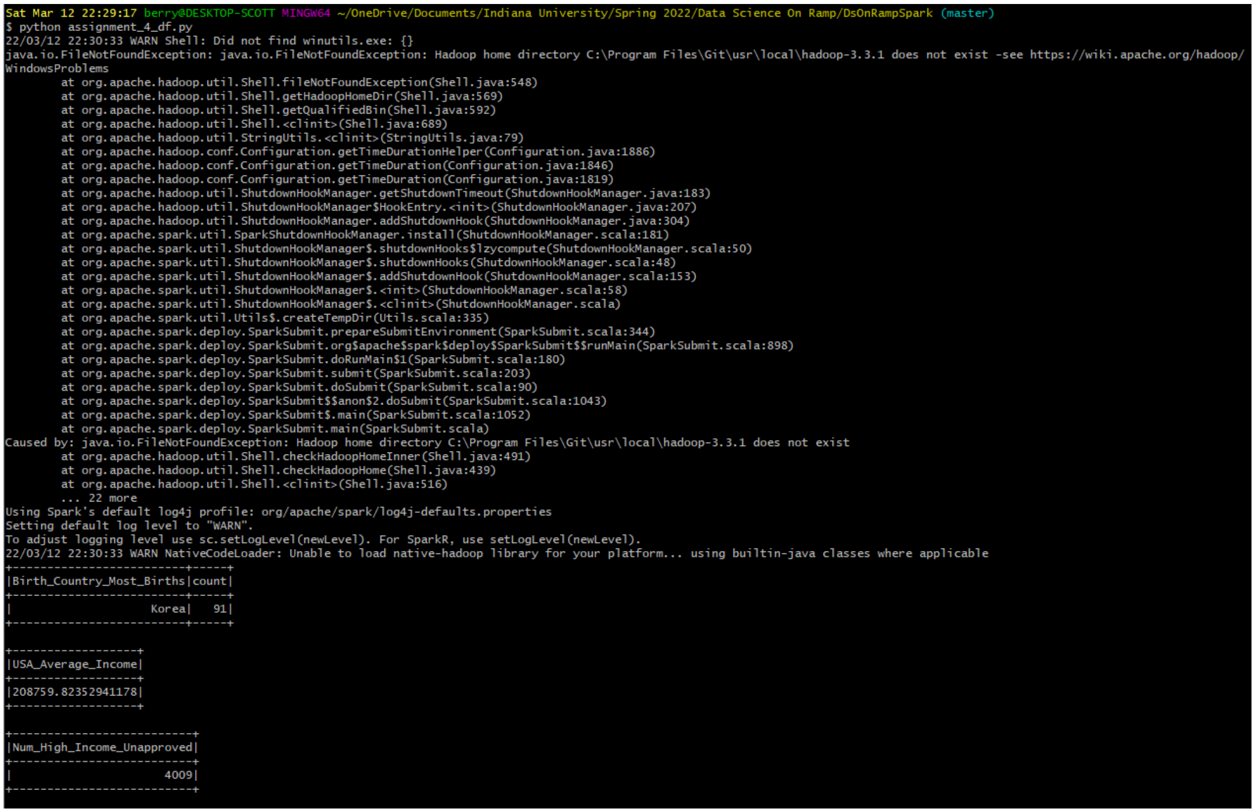
Scott Berry

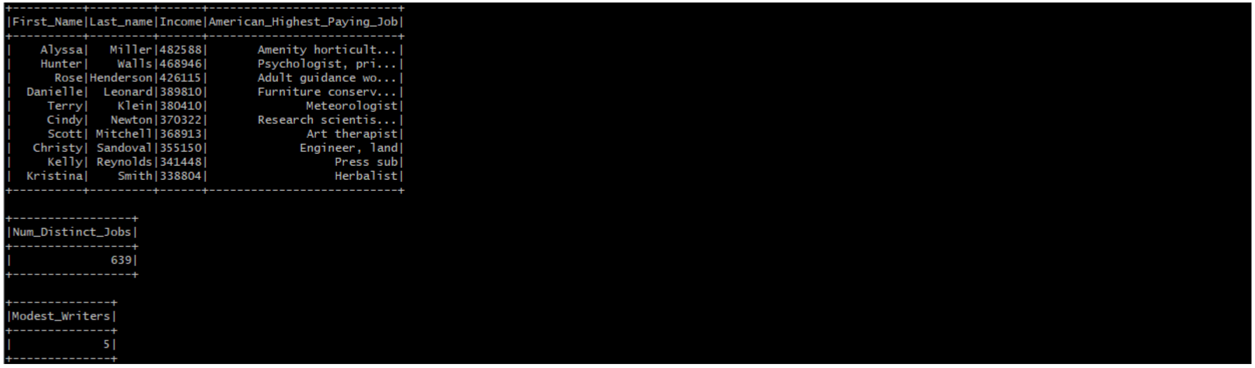
1) Transformation and Actions



```
Sat Mar 12 22:29:17 berry@DESKTOP-SCOTT MINGW64 ~/OneDrive/Documents/Indiana University/Spring 2022/Data Science On Ramp/DsOnRampSpark (master)
$ python assignment_4_df.py
22/03/12 22:30:33 WARN Shell: Did not find winutils.exe: {}
java.io.FileNotFoundException: java.io.FileNotFoundException: Hadoop home directory C:\Program Files\Git\usr\local\hadoop-3.3.1 does not exist -see https://wiki.apache.org/hadoop/
WindowsProblems
        at org.apache.hadoop.util.Shell.fileNotFoundException(Shell.java:548)
        at org.apache.hadoop.util.Shell.getHadoopHomeDir(Shell.java:569)
        at org.apache.hadoop.util.Shell.getQualifiedBin(Shell.java:592)
        at org.apache.hadoop.util.Shell.<clinit>(Shell.java:689)
        at org.apache.hadoop.util.StringUtils.<clinit>(StringUtils.java:79)
        at org.apache.hadoop.conf.Configuration.getTimeDurationHelper(Configuration.java:1886)
        at org.apache.hadoop.conf.Configuration.getTimeDuration(Configuration.java:1846)
        at org.apache.hadoop.conf.Configuration.getTimeDuration(Configuration.java:1819)
        at org.apache.hadoop.util.ShutdownHookManager.getShutdownTimeout(ShutdownHookManager.java:183)
        at org.apache.hadoop.util.ShutdownHookManager$HookEntry.<init>(ShutdownHookManager.java:207)
        at org.apache.hadoop.util.ShutdownHookManager.addShutdownHook(ShutdownHookManager.java:304)
        at org.apache.spark.util.SparkShutdownHookManager.install(ShutdownHookManager.scala:181)
        at org.apache.spark.util.ShutdownHookManager$.shutdownHooks$lzycompute(ShutdownHookManager.scala:50)
        at org.apache.spark.util.ShutdownHookManager$.shutdownHooks(ShutdownHookManager.scala:48)
        at org.apache.spark.util.ShutdownHookManager$.addShutdownHook(ShutdownHookManager.scala:153)
        at org.apache.spark.util.ShutdownHookManager$.<init>(ShutdownHookManager.scala:58)
        at org.apache.spark.util.ShutdownHookManager$.<clinit>(ShutdownHookManager.scala)
        at org.apache.spark.util.Utils$.createTempDir(Utils.scala:335)
        at org.apache.spark.deploy.SparkSubmit.prepareSubmitEnvironment(SparkSubmit.scala:344)
        at org.apache.spark.deploy.SparkSubmit.org$apache$spark$deploy$SparkSubmit$$runMain(SparkSubmit.scala:898)
        at org.apache.spark.deploy.SparkSubmit.doRunMain$1(SparkSubmit.scala:180)
        at org.apache.spark.deploy.SparkSubmit.submit(SparkSubmit.scala:203)
        at org.apache.spark.deploy.SparkSubmit.doSubmit(SparkSubmit.scala:90)
        at org.apache.spark.deploy.SparkSubmit$$anon$2.doSubmit(SparkSubmit.scala:1043)
        at org.apache.spark.deploy.SparkSubmit$.main(SparkSubmit.scala:1052)
        at org.apache.spark.deploy.SparkSubmit.main(SparkSubmit.scala)
Caused by: java.io.FileNotFoundException: Hadoop home directory C:\Program Files\Git\usr\local\hadoop-3.3.1 does not exist
        at org.apache.hadoop.util.Shell.checkHadoopHomeInner(Shell.java:491)
        at org.apache.hadoop.util.Shell.checkHadoopHome(Shell.java:439)
        at org.apache.hadoop.util.Shell.<clinit>(Shell.java:516)
        ... 22 more
Using Spark's default log4j profile: org/apache/spark/log4j-defaults.properties
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
22/03/12 22:30:33 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
+-------------------------+-----+
|Birth_Country_Most_Births|count|
+-------------------------+-----+
|                    Korea|   91|
+-------------------------+-----+

+------------------+
|USA_Average_Income|
+------------------+
|208759.82352941178|
+------------------+

+--------------------------+
|Num_High_Income_Unapproved|
+--------------------------+
|                      4009|
+--------------------------+
```

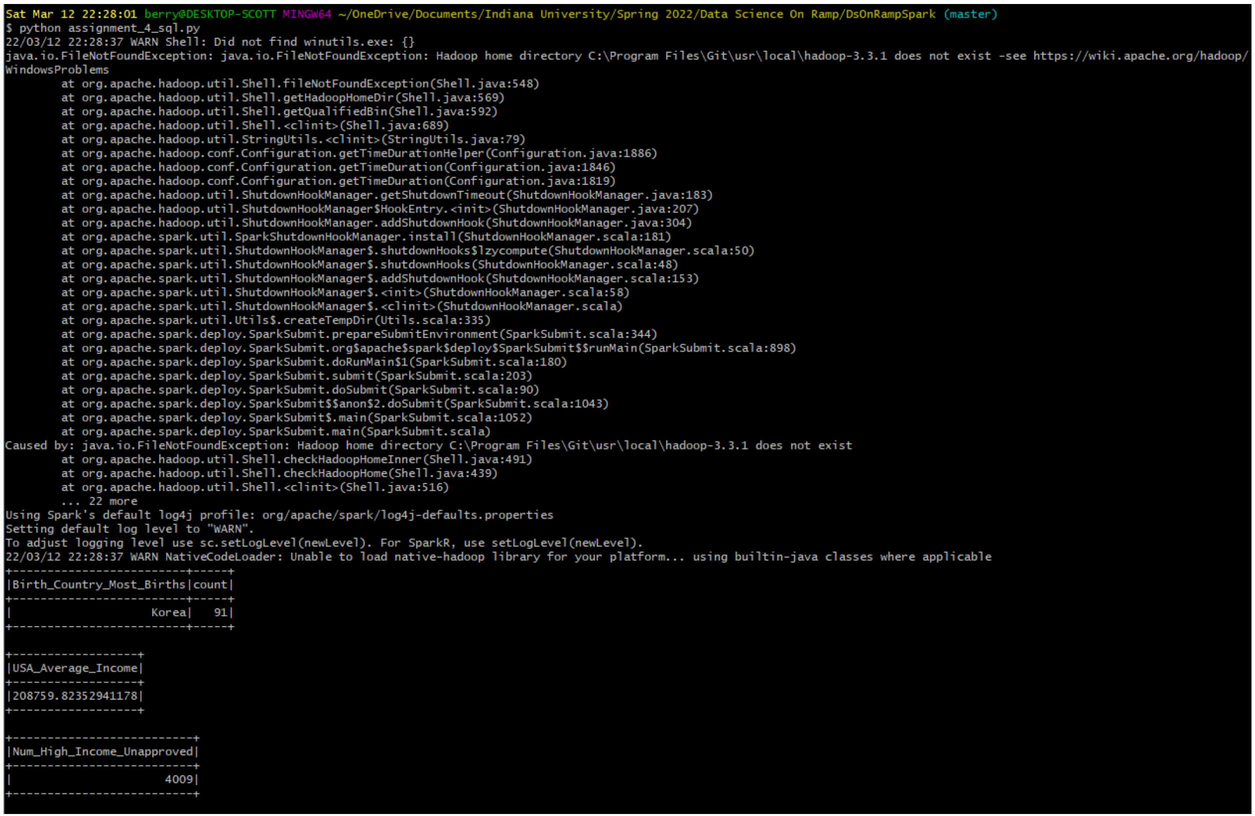


```
+----------+---------+------+---------------------------+
|First_Name|Last_name|Income|American_Highest_Paying_Job|
+----------+---------+------+---------------------------+
|    Alyssa|   Miller|482588|       Amenity horticult...|
|    Hunter|    Walls|468946|       Psychologist, pri...|
|      Rose|Henderson|426115|       Adult guidance wo...|
|  Danielle|  Leonard|389810|       Furniture conserv...|
|     Terry|    Klein|380410|              Meteorologist|
|     Cindy|   Newton|370322|       Research scientis...|
|     Scott| Mitchell|368913|              Art therapist|
|   Christy| Sandoval|355150|             Engineer, land|
|     Kelly| Reynolds|341448|                  Press sub|
|  Kristina|    Smith|338804|                  Herbalist|
+----------+---------+------+---------------------------+

+-----------------+
|Num_Distinct_Jobs|
+-----------------+
|              639|
+-----------------+

+--------------+
|Modest_Writers|
+--------------+
|             5|
+--------------+
```

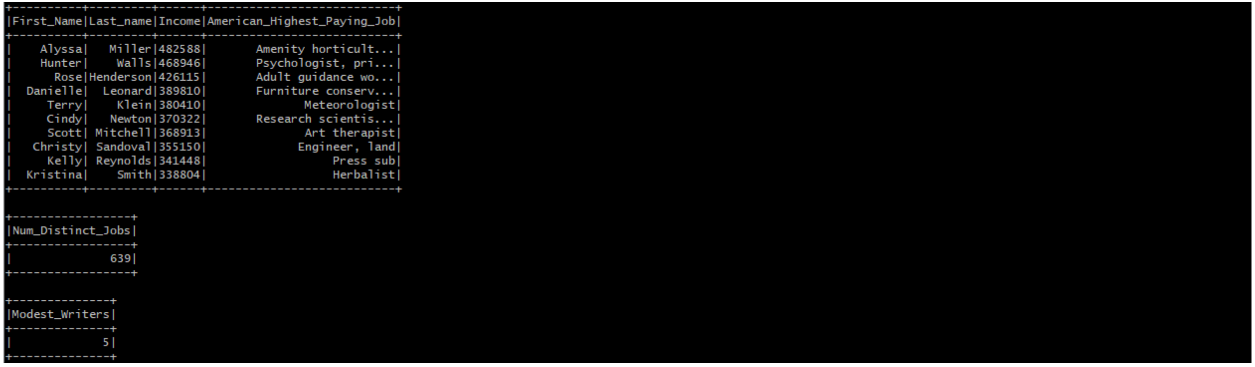
2) Spark SQL



```
Sat Mar 12 22:28:01 berry@DESKTOP-SCOTT MINGW64 ~/OneDrive/Documents/Indiana University/Spring 2022/Data Science On Ramp/DsOnRampSpark (master)
$ python assignment_4_sql.py
22/03/12 22:28:37 WARN Shell: Did not find winutils.exe: {}
java.io.FileNotFoundException: java.io.FileNotFoundException: Hadoop home directory C:\Program Files\Git\usr\local\hadoop-3.3.1 does not exist -see https://wiki.apache.org/hadoop/
WindowsProblems
        at org.apache.hadoop.util.Shell.fileNotFoundException(Shell.java:548)
        at org.apache.hadoop.util.Shell.getHadoopHomeDir(Shell.java:569)
        at org.apache.hadoop.util.Shell.getQualifiedBin(Shell.java:592)
        at org.apache.hadoop.util.Shell.<clinit>(Shell.java:689)
        at org.apache.hadoop.util.StringUtils.<clinit>(StringUtils.java:79)
        at org.apache.hadoop.conf.Configuration.getTimeDurationHelper(Configuration.java:1886)
        at org.apache.hadoop.conf.Configuration.getTimeDuration(Configuration.java:1846)
        at org.apache.hadoop.conf.Configuration.getTimeDuration(Configuration.java:1819)
        at org.apache.hadoop.util.ShutdownHookManager.getShutdownTimeout(ShutdownHookManager.java:183)
        at org.apache.hadoop.util.ShutdownHookManager$HookEntry.<init>(ShutdownHookManager.java:207)
        at org.apache.hadoop.util.ShutdownHookManager.addShutdownHook(ShutdownHookManager.java:304)
        at org.apache.spark.util.SparkShutdownHookManager.install(ShutdownHookManager.scala:181)
        at org.apache.spark.util.ShutdownHookManager$.shutdownHooks$lzycompute(ShutdownHookManager.scala:50)
        at org.apache.spark.util.ShutdownHookManager$.shutdownHooks(ShutdownHookManager.scala:48)
        at org.apache.spark.util.ShutdownHookManager$.addShutdownHook(ShutdownHookManager.scala:153)
        at org.apache.spark.util.ShutdownHookManager$.<init>(ShutdownHookManager.scala:58)
        at org.apache.spark.util.ShutdownHookManager$.<clinit>(ShutdownHookManager.scala)
        at org.apache.spark.util.Utils$.createTempDir(Utils.scala:335)
        at org.apache.spark.deploy.SparkSubmit.prepareSubmitEnvironment(SparkSubmit.scala:344)
        at org.apache.spark.deploy.SparkSubmit.org$apache$spark$deploy$SparkSubmit$$runMain(SparkSubmit.scala:898)
        at org.apache.spark.deploy.SparkSubmit.doRunMain$1(SparkSubmit.scala:180)
        at org.apache.spark.deploy.SparkSubmit.submit(SparkSubmit.scala:203)
        at org.apache.spark.deploy.SparkSubmit.doSubmit(SparkSubmit.scala:90)
        at org.apache.spark.deploy.SparkSubmit$$anon$2.doSubmit(SparkSubmit.scala:1043)
        at org.apache.spark.deploy.SparkSubmit$.main(SparkSubmit.scala:1052)
        at org.apache.spark.deploy.SparkSubmit.main(SparkSubmit.scala)
Caused by: java.io.FileNotFoundException: Hadoop home directory C:\Program Files\Git\usr\local\hadoop-3.3.1 does not exist
        at org.apache.hadoop.util.Shell.checkHadoopHomeInner(Shell.java:491)
        at org.apache.hadoop.util.Shell.checkHadoopHome(Shell.java:439)
        at org.apache.hadoop.util.Shell.<clinit>(Shell.java:516)
        ... 22 more
Using Spark's default log4j profile: org/apache/spark/log4j-defaults.properties
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
22/03/12 22:28:37 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
+------------------------+-----+
|Birth_Country_Most_Births|count|
+------------------------+-----+
|                   Korea|   91|
+------------------------+-----+

+------------------+
|USA_Average_Income|
+------------------+
|208759.82352941178|
+------------------+

+--------------------------+
|Num_High_Income_Unapproved|
+--------------------------+
|                      4009|
+--------------------------+
```

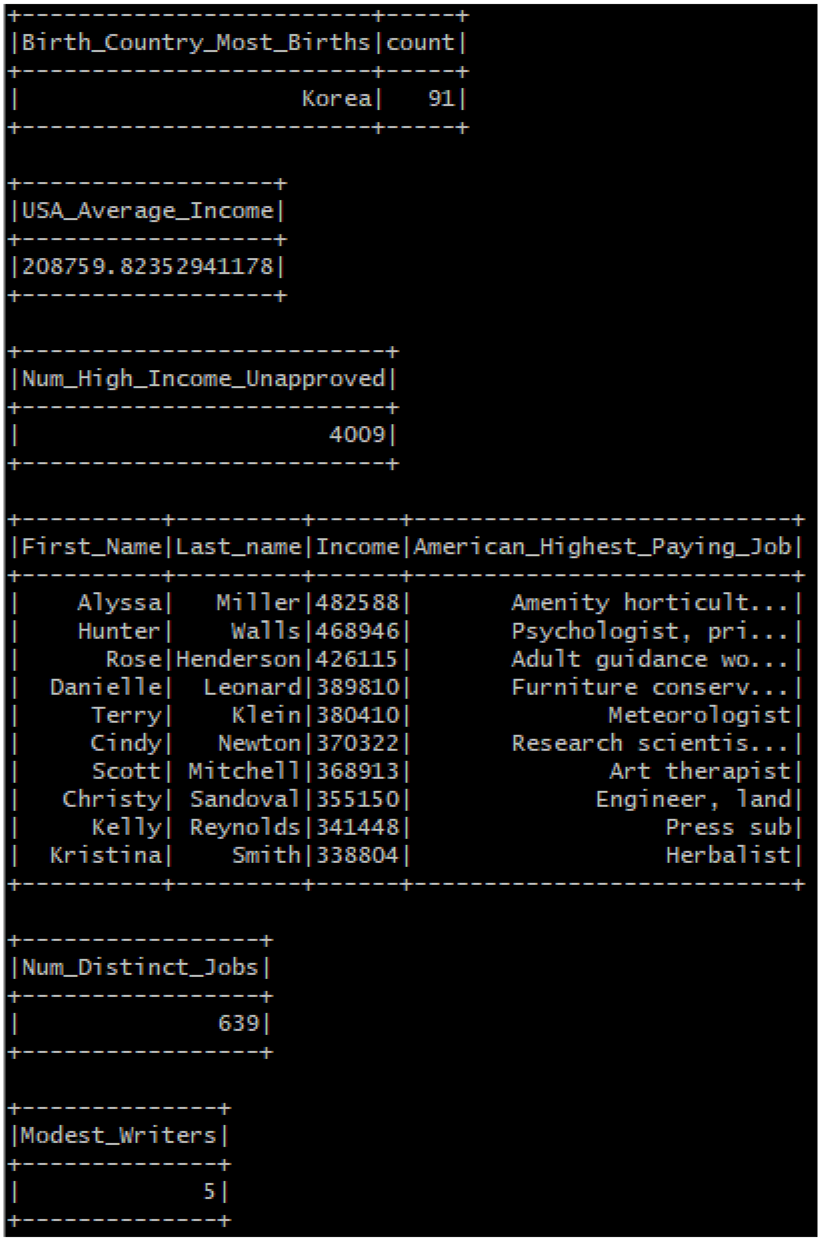


```
+----------+---------+------+----------------------------+
|First_Name|Last_name|Income|American_Highest_Paying_Job|
+----------+---------+------+----------------------------+
|    Alyssa|   Miller|482588|        Amenity horticult...|
|    Hunter|    Walls|468946|        Psychologist, pri...|
|      Rose|Henderson|426115|        Adult guidance wo...|
|  Danielle|  Leonard|389810|        Furniture conserv...|
|     Terry|    Klein|380410|               Meteorologist|
|     Cindy|   Newton|370322|        Research scientis...|
|     Scott| Mitchell|368913|               Art therapist|
|   Christy| Sandoval|355150|              Engineer, land|
|     Kelly| Reynolds|341448|                   Press sub|
|  Kristina|    Smith|338804|                   Herbalist|
+----------+---------+------+----------------------------+

+-----------------+
|Num_Distinct_Jobs|
+-----------------+
|              639|
+-----------------+

+--------------+
|Modest_Writers|
+--------------+
|             5|
+--------------+
```

This view of the results is more legible but doesn’t include full logging (results were same for both)



```
+-------------------------+-----+
|Birth_Country_Most_Births|count|
+-------------------------+-----+
|                    Korea|   91|
+-------------------------+-----+

+------------------+
|USA_Average_Income|
+------------------+
|208759.82352941178|
+------------------+

+--------------------------+
|Num_High_Income_Unapproved|
+--------------------------+
|                      4009|
+--------------------------+

+----------+---------+------+---------------------------+
|First_Name|Last_name|Income|American_Highest_Paying_Job|
+----------+---------+------+---------------------------+
|    Alyssa|   Miller|482588|       Amenity horticult...|
|    Hunter|    Walls|468946|       Psychologist, pri...|
|      Rose|Henderson|426115|       Adult guidance wo...|
|  Danielle|  Leonard|389810|       Furniture conserv...|
|     Terry|    Klein|380410|              Meteorologist|
|     Cindy|   Newton|370322|       Research scientis...|
|     Scott| Mitchell|368913|              Art therapist|
|   Christy| Sandoval|355150|             Engineer, land|
|     Kelly| Reynolds|341448|                  Press sub|
|  Kristina|    Smith|338804|                  Herbalist|
+----------+---------+------+---------------------------+

+-----------------+
|Num_Distinct_Jobs|
+-----------------+
|              639|
+-----------------+

+--------------+
|Modest_Writers|
+--------------+
|             5|
+--------------+
```