Data Analysis Assessment

We would be analyzing United States Census Bureau's 2017 Basic Monthly CPS, using Apache Spark (with preferably Python or Scala).

Thoroughly read and follow the instructions for the task below. At the end, kindly share your code and snapshots of the data output via github.com:

Step 1: Visit the United States Census Bureau website and access the 2017 Basic Monthly CPS page.

Step 2: Download the DOS/Windows for December zip file and extract the *dat* file. It contains data captured per respondent.

Step 3: Using the data dictionary file, extract the following information, and show a sample of 10 records only:

1. Full household identifier.
2. Time of interview in *YYYY/MMM* format.
3. Final outcome of the survey.
4. Type of housing unit.
5. Household type.
6. Apartment/Household has a telephone.
7. Apartment/Household can access a telephone elsewhere.
8. Is telephone interview acceptable for the responder.
9. Type of interview.
10. Family income range.
11. Geographical division/location.
12. Race.

Step 4: Using the custom data extracted in step 3, answer the below questions:

1. What is the count of responders per family income range (show all)?
2. What is the count of responders per geographical division/location and race (show top 10)?
3. How many responders do not have telephone in their house, but can access a telephone elsewhere and telephone interview is accepted?
4. How many responders can access a telephone, but telephone interview is not accepted?

Note: Where a value is encoded, always return the actual (decoded) value.