1. Data extraction
   1. Import full national dataset for 2019: 17 million rows, 99 columns
   2. Subset the full national dataset by state = TN: 388K rows, 99 columns
   3. Export the TN dataset to .csv
   4. Write an import script
2. Data cleaning
   1. Identify the dependent variable(s) that contain information about mortgage application approval or denial (use the HMDA data dictionary)
      1. action\_taken
      2. denial\_reason\_1
      3. denial\_reason\_2
      4. denial\_reason\_3
      5. denial\_reason\_4
   2. Identify the independent variables (use the HMDA data dictionary)
      1. Identify the continuous variables
         1. income
         2. property\_value
         3. interest\_rate
      2. Identify the categorical variables (use the HMDA data dictionary)
         1. activity\_year
         2. lei
         3. county\_code
         4. census\_tract
         5. …
   3. Remove variables that have the same value for all rows (use describe() to see leveland frequencies)
      1. state\_code
      2. activity\_year
   4. Remove variables with duplicated information
      1. applicant\_age\_above\_62
      2. co\_applicant\_age\_above\_62
   5. Remove variables with all missing or mostly missing values (use contents() to see NAs)
      1. total\_points\_and\_fees
      2. discount\_points
      3. lender\_credits
      4. prepayment\_penalty\_term
      5. intro\_rate\_period
      6. multifamily\_affordable\_units
      7. applicant\_race\_2
      8. applicant\_race\_3
      9. applicant\_race\_4
      10. applicant\_race\_5
      11. co\_applicant\_race\_2
      12. co\_applicant\_race\_3
      13. co\_applicant\_race\_4
      14. co\_applicant\_race\_5
      15. applicant\_ethnicity\_2
      16. applicant\_ethnicity\_3
      17. applicant\_ethnicity\_4
      18. applicant\_ethnicity\_5
      19. co\_applicant\_ethnicity\_2
      20. co\_applicant\_ethnicity\_3
      21. co\_applicant\_ethnicity\_4
      22. co\_applicant\_ethnicity\_5
      23. aus\_2
      24. aus\_3
      25. aus\_4
      26. aus\_5
      27. denial\_reason\_2
      28. denial\_reason\_3
      29. denial\_reason\_4
   6. Examine each variable to understand the information and levels
      1. ‘derived’ columns - do the collapsed columns make sense or lose data?
         1. derived\_msa\_md – combines state, county, census tract; keep as this variable has fewer levels than county & census tract
         2. derived\_loan\_product\_type - combines
         3. derived\_dwelling\_category – combines
         4. derived\_race – combines race designation of primary and co-applicant
         5. derived\_ethnicity – combines ethnicity designation of primary and co-applicant
         6. derived\_sex – we lose information about the sex of the primary applicant when there is a co-applicant; “Joint” does not indicate the sex of the primary applicant; will need to address this or not use this column
      2. The co-applicant fields will have a value of ‘Not applicable’ if there is no co-applicant, which may skew the results
      3. continuous variables have some “Exempt” or “Not applicable” if the data reported is for a mortgage that was purchased or the applicant was “not a natural person”, or if the lending institution is not required to report that data.
      4. factor variables have some “Exempt” or “Not applicable” if the data reported is for a mortgage that was purchased or the applicant was “not a natural person”, or if the lending institution is not required to report that data.
   7. Examine each