```
/*******************************
 * NAME: KELLY ADU - GYAMFI     *
 * STA 551 PROJECT 1            *
 * PURPOSE: INTEGRATING DATA    *
 * DUE DATE: SEPTEMBER 17, 2025 *
*******************************/



/*IMPORTING ALL 4 DATASETS*/

PROC IMPORT DATAFILE =
"\\apporto.com\dfs\WCUPA\Users\1052757_wcupa\Desktop\551\LECTURE
3\EDUCATION.CSV"
OUT = EDUCATION_RAW
DBMS = CSV
REPLACE;
GETNAMES = YES;
GUESSINGROWS = MAX;
RUN;

PROC IMPORT DATAFILE =
"\\apporto.com\dfs\WCUPA\Users\1052757_wcupa\Desktop\551\LECTURE
3\countypresidential_election_2000-2020.CSV"
OUT = ELECTION_DATA_RAW
DBMS = CSV
REPLACE;
GETNAMES = YES;
RUN;

PROC IMPORT DATAFILE =
"\\apporto.com\dfs\WCUPA\Users\1052757_wcupa\Desktop\551\LECTURE
3\PovertyEstimates.CSV"
OUT = POVERTY_RAW
DBMS = CSV
REPLACE;
GETNAMES = YES;
RUN;

PROC IMPORT DATAFILE =
"\\apporto.com\dfs\WCUPA\Users\1052757_wcupa\Desktop\551\LECTURE
3\UNEMPLOYMENT.CSV"
OUT = UNEMPLOYMENT_RAW
DBMS = CSV
REPLACE;
GETNAMES = YES;
RUN;



/*============================================================
   STEP 1: CLEAN PRESIDENTIAL ELECTION DATA
   - Keep only 2020 election results
   - Retain Democrat and Republican votes
   - Identify the winning party and total votes per county
============================================================*/
PROC SQL;
```

```
CREATE TABLE ELECTION_2020 AS
SELECT  COUNTY_FIPS,
        STATE,
        COUNTY_NAME AS COUNTY,
        PARTY,
        CANDIDATEVOTES
FROM ELECTION_DATA_RAW
WHERE YEAR = 2020
AND (PARTY="DEMOCRAT" OR PARTY="REPUBLICAN");
QUIT;

/* Aggregate total votes by county and party */
PROC SQL;
CREATE TABLE ELECTION_VOTES AS
SELECT COUNTY_FIPS,
       STATE,
         COUNTY,
         PARTY AS WINNING_PARTY,
         SUM(CANDIDATEVOTES) AS TOTAL_VOTES
FROM ELECTION_2020
GROUP BY COUNTY_FIPS, STATE, COUNTY, PARTY;
QUIT;

/* Keep only the winning party per county (highest total votes) */
PROC SQL;
CREATE TABLE ELECTION_CLEAN AS
SELECT A.COUNTY_FIPS,
       A.STATE,
         A.COUNTY,
         A.WINNING_PARTY,
         A.TOTAL_VOTES
FROM ELECTION_VOTES AS A
WHERE A.TOTAL_VOTES = (SELECT MAX(B.TOTAL_VOTES)
FROM ELECTION_VOTES AS B
WHERE A.COUNTY_FIPS = B.COUNTY_FIPS);
QUIT;


/*============================================================
   STEP 2: CLEAN UNEMPLOYMENT DATA
   - Parse indicator and year from raw data
   - Keep only 2020 unemployment rate per county
============================================================*/
PROC SQL;
CREATE TABLE UNEMPLOYED_CLEAN AS
SELECT FIPS_CODE AS COUNTY_FIPS,
       STATE,
         AREA_NAME,
         SCAN(ATTRIBUTE, 1,'_') AS INDICATOR,
         INPUT(SCAN(ATTRIBUTE, -1, '_') , 4.) AS YEAR,
         VALUE
      FROM UNEMPLOYMENT_RAW;
QUIT;

/* Keep only unemployment rate for 2020 */
PROC SQL;
CREATE TABLE UNEMPLOYMENT_RATE_2020 AS
```

```
SELECT COUNTY_FIPS,
       STATE,
         AREA_NAME,
         VALUE AS UNWMPLOYMENT_RATE
  FROM UNEMPLOYED_CLEAN
  WHERE UPCASE(INDICATOR) = "UNEMPLOYMENT"
  AND YEAR = 2020 ;
  QUIT;

/*============================================================
   STEP 3: CLEAN POVERTY DATA
   - Extract indicator and year
   - Keep only 2019 poverty rate (PCTPOVALL_2019)
============================================================*/
PROC SQL;
CREATE TABLE POVERTY_CLEAN AS
SELECT FIPSTXT AS COUNTY_FIPS,
       STABR AS STATE,
         AREA_NAME,
         SUBSTR(ATTRIBUTE, 1, FIND(ATTRIBUTE, '_', -LENGTH(ATTRIBUTE)) -1) AS
INDICATOR,
         SUBSTR(ATTRIBUTE, FIND(ATTRIBUTE, '_', -LENGTH(ATTRIBUTE)) +1) AS
YEAR,
         VALUE
FROM POVERTY_RAW;
QUIT;

/* Keep poverty rate for 2019 */
PROC SQL;
CREATE TABLE POVERTY_2019 AS
SELECT COUNTY_FIPS,
       STATE,
         AREA_NAME,
         VALUE AS PCTPOVALL_2019
  FROM POVERTY_CLEAN
  WHERE UPCASE(INDICATOR) = "PCTPOVALL"
  AND INPUT(YEAR,4.) = 2019;
  QUIT;

/*============================================================
   STEP 4: CLEAN EDUCATION DATA
   - Keep education levels for 2015-2019
   - Rename variables for clarity
============================================================*/
PROC SQL;
CREATE TABLE EDUCATION_RENAMED AS
SELECT FIPS_CODE AS COUNTY_FIPS,
       STATE,
         AREA_NAME,
       VAR40 AS LESS_THAN_HS,
       High_school_diploma_only__2015_ AS HS_ONLY ,
         VAR42 AS SOME_COLLEGE_15_19,
         VAR43 AS BACHELORS_15_19,
       VAR44 AS PCT_LESS_THAN_HS_15_19,
         VAR45 AS PCT_HS_ONLY_15_19,
         VAR46 AS PCT_SOME_COLLEGE_15_19,
         VAR47 AS PCT_BAC_0R_HIGH_15_19
```

```
FROM EDUCATION_RAW;
QUIT;


/*STEP 5:MERGING ALL THE FOUR DATASETS*/
PROC SQL;
    CREATE TABLE INTEGRATED_DATA AS
    SELECT EDU.COUNTY_FIPS,
           EDU.STATE,
           EDU.AREA_NAME,
           /* EDUCATION VARIABLES*/
           EDU.LESS_THAN_HS,
           EDU.HS_ONLY,
           EDU.SOME_COLLEGE_15_19,
           EDU.BACHELORS_15_19,
           EDU.PCT_LESS_THAN_HS_15_19,
           EDU.PCT_HS_ONLY_15_19,
           EDU.PCT_SOME_COLLEGE_15_19,
           EDU.PCT_BAC_0R_HIGH_15_19,
           /* ELECTION VARIABLES*/
           EL.WINNING_PARTY,
           EL.TOTAL_VOTES,
           /* UNEMPLOYMENT */
           UN.UNWMPLOYMENT_RATE,
           /* POVERTY */
           PV.PCTPOVALL_2019
    FROM EDUCATION_RENAMED AS EDU
    LEFT JOIN ELECTION_CLEAN AS EL
        ON EDU.COUNTY_FIPS = EL.COUNTY_FIPS
    LEFT JOIN UNEMPLOYMENT_RATE_2020 AS UN
        ON EDU.COUNTY_FIPS = UN.COUNTY_FIPS
    LEFT JOIN POVERTY_2019 AS PV
        ON EDU.COUNTY_FIPS = PV.COUNTY_FIPS;
QUIT;

/*STEP 6 : EXPORT FINAL DATA TO CSV*/
PROC EXPORT DATA = INTEGRATED_DATA
OUTFILE = "\\apporto.com\dfs\WCUPA\Users\1052757_wcupa\Desktop\551\LECTURE
3\INTEGRATED_DATA.CSV"
DBMS = CSV
REPLACE;
RUN;



/*------------------------------------------------------------
  Summary of Integrated Dataset

  The integrated dataset was created by combining four separate
  data sources: presidential election results (2000-2020),
  unemployment data, poverty estimates, and education statistics.
  The data were filtered, cleaned, and merged by county using
  the FIPS code as the unique identifier.

  - Presidential Election Data:
      • Restricted to the year 2020 and the two major parties
        (Democrat, Republican).
```

```
      • Retains the winning party and total votes received by
        that party for each county.
  - Unemployment Data:
      • Limited to the 2020 unemployment rate (or most recent).
  - Poverty Data:
      • Includes only the 2019 poverty rate (PCTPOVALL_2019).
  - Education Data:
      • Includes percentages of residents with less than high
        school, high school diploma only, some college, and
        bachelor's degree or higher (2015–2019).

  Final Dataset:
      • Approx. 3,100 counties
      • 11 variables
      • One record per county

  Variables in Final Dataset:
      - COUNTY_FIPS        (Numeric: County identifier)
      - STATE              (Character: State name)
      - COUNTY             (Character: County name)
      - TOTAL_VOTES        (Numeric: Total votes of winning party)
      - WINNING_PARTY      (Character: Democrat or Republican)
      - UNEMPLOYMENT_RATE  (Numeric: 2020 unemployment rate)
      - POVERTY_RATE       (Numeric: 2019 poverty rate)
      - LESS_HS            (Numeric: % less than HS diploma, 2015–2019)
      - HS_DIPLOMA         (Numeric: % HS diploma only, 2015–2019)
      - SOME_COLLEGE       (Numeric: % some college, 2015–2019)
      - BACHELOR_OR_HIGHER (Numeric: % bachelor's degree or higher, 2015–
2019)

  This dataset provides a structured foundation for exploratory
  data analysis (EDA), including the study of relationships
  between socioeconomic conditions and presidential election
  outcomes.
-------------------------------------------------------------*/
```