Katie Mowry

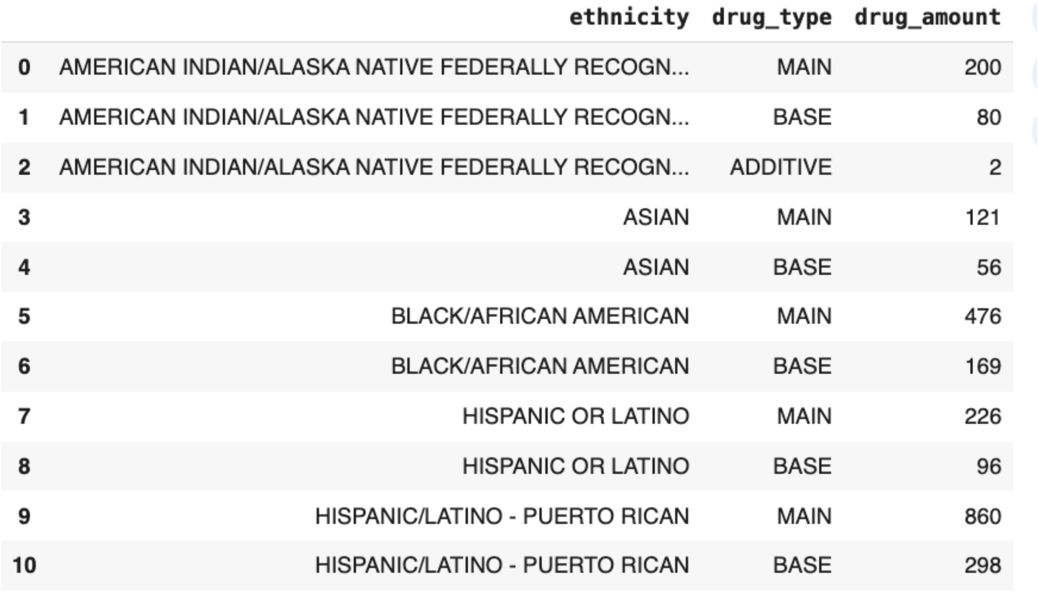
DATA ENG 300 Homework 2 Responses

Part I

**All SQL queries are within the DATA ENG 300 Homework 2 Part I.inpyb file

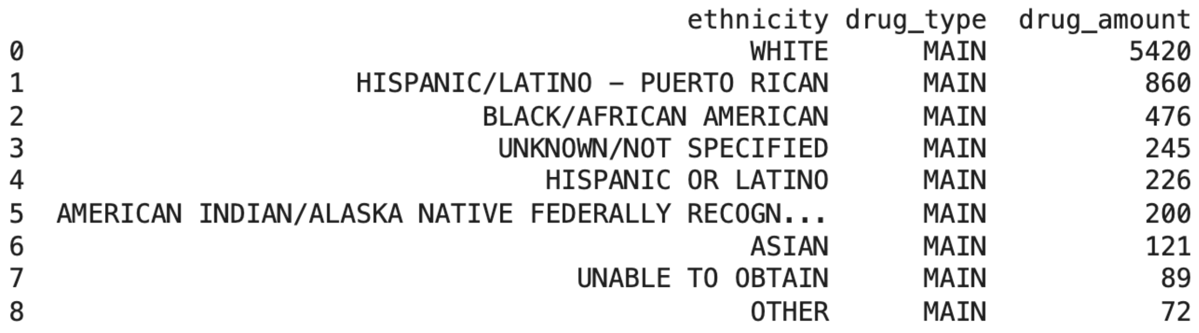
1b. The SQL query selects ethnicity from the ADMISSIONS table and drug_type from the PRESCRPTIONS table, then counts the number of prescriptions for each ethnicity and drug_type combination as drug_amount. The tables are joined using hadm_id to ensure that each prescription is correctly associated with the hospital admission since a single patient (subject_id) can have multiple hospital admissions. The results are grouped by ethnicity and drug_type to count the total number of drug_amount and ordered in descending order to identify the top used drug type per group. This displays all combinations of ethnicity and drug_type, along with how many times each drug type was prescribed for each ethnicity. To identify the top usage in each ethnicity group, the top drug_type for each ethnicity is printed using Python and placed in descending order. This resulting table is shown below.

1c. The following table shows the first few lines of the dataframe from the SQL query, showing the type of dugs and their total amount used by ethnicity.



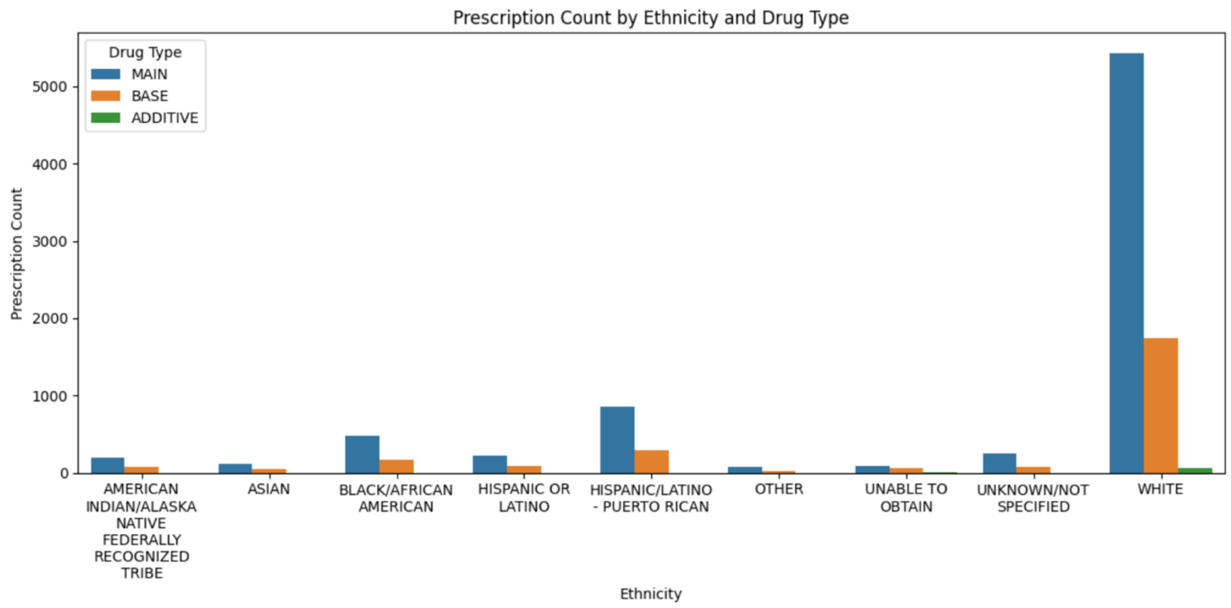
	ethnicity	drug_type	drug_amount
0	AMERICAN INDIAN/ALASKA NATIVE FEDERALLY RECOGN...	MAIN	200
1	AMERICAN INDIAN/ALASKA NATIVE FEDERALLY RECOGN...	BASE	80
2	AMERICAN INDIAN/ALASKA NATIVE FEDERALLY RECOGN...	ADDITIVE	2
3	ASIAN	MAIN	121
4	ASIAN	BASE	56
5	BLACK/AFRICAN AMERICAN	MAIN	476
6	BLACK/AFRICAN AMERICAN	BASE	169
7	HISPANIC OR LATINO	MAIN	226
8	HISPANIC OR LATINO	BASE	96
9	HISPANIC/LATINO - PUERTO RICAN	MAIN	860
10	HISPANIC/LATINO - PUERTO RICAN	BASE	298

The following tables shows the top usage in each ethnicity group after grouping the results using Python.



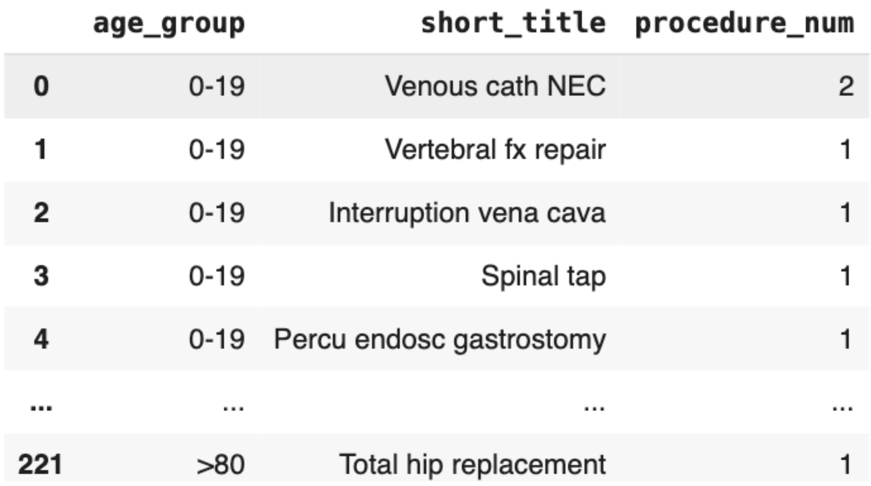
```
                                           ethnicity drug_type  drug_amount
0                                              WHITE      MAIN         5420
1                     HISPANIC/LATINO - PUERTO RICAN      MAIN          860
2                             BLACK/AFRICAN AMERICAN      MAIN          476
3                             UNKNOWN/NOT SPECIFIED       MAIN          245
4                                HISPANIC OR LATINO       MAIN          226
5  AMERICAN INDIAN/ALASKA NATIVE FEDERALLY RECOGN...      MAIN          200
6                                              ASIAN      MAIN          121
7                                   UNABLE TO OBTAIN      MAIN           89
8                                              OTHER      MAIN           72
```

1d. Among the three drug types (MAIN, BASE, and ADDITIVE), MAIN drugs are the most used across all ethnic groups. White patients have the highest overall prescription count for all drugs, and particularly for MAIN drugs. Hispanic/Latino – Puerto Rican and Black/African American patients have the next highest MAIN drug counts. A bar chart was created to visualize the distribution of drug usage for each drug type by ethnicity. It can be seen that there is a large disparity in drug usage between ethnicities, especially for MAIN drugs, with white patients receiving a significantly higher amount.



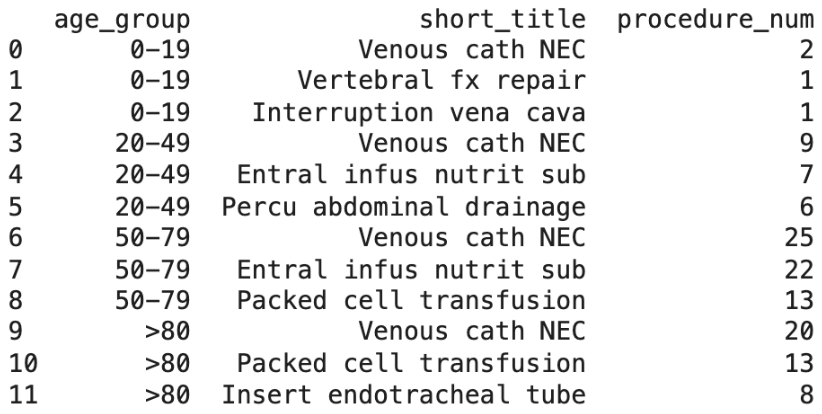
2b. First, the PATIENTS and ADMISSIONS tables are joined on subject_id to link the patient's gender to admittime. Then, PROCEDURES_ICD and D_ADMMISSIONS tables are joined using hadm_id to associate each admission with the correct procedure, and PROCEUDRES_ICD and D_ICD_PROCEDURES are joined using the icd9_code, which is the code specific to the procedure type. The difference between the admitted time in the ADMISSIONS table and the date of birth from the PATIENTS table was calculated to find the age that the patient was when the procedure was performed. The results were grouped by age (0-19, 20-49, 50-79, and >80), and the number of times each procedure was performed within each age group was counted. The final output is sorted by age group and number of procedures. Python was then used to display the top three procedures, along with the name of the procedures, performed in each age group.

2c. The following table shows the top few lines of the dataframe from the SQL query, showing the first few procedures, age groups, and procedure numbers.



	age_group	short_title	procedure_num
0	0-19	Venous cath NEC	2
1	0-19	Vertebral fx repair	1
2	0-19	Interruption vena cava	1
3	0-19	Spinal tap	1
4	0-19	Percu endosc gastrostomy	1
...	...	...	...
221	>80	Total hip replacement	1

The following table shows the top three procedures, along with the name of the procedures, performed in each age group after grouping the results using Python.

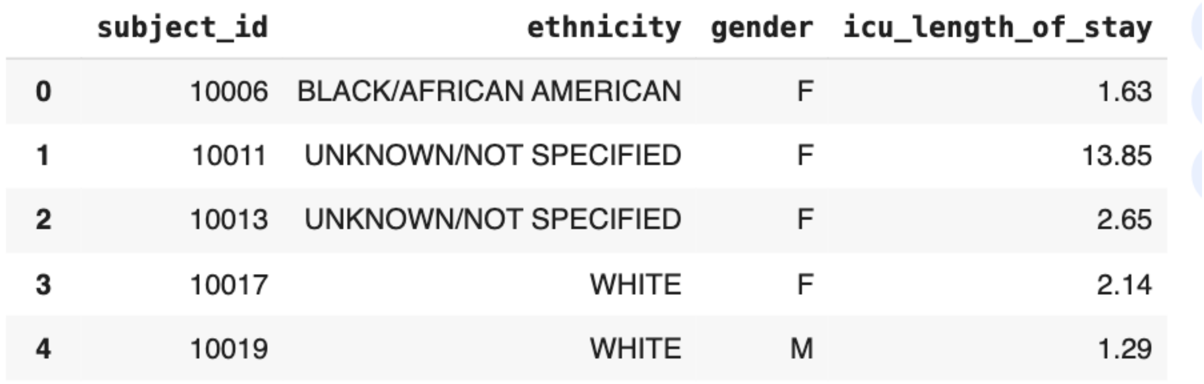


```
   age_group               short_title  procedure_num
0       0–19           Venous cath NEC              2
1       0–19       Vertebral fx repair              1
2       0–19    Interruption vena cava              1
3      20–49           Venous cath NEC              9
4      20–49    Entral infus nutrit sub              7
5      20–49  Percu abdominal drainage              6
6      50–79           Venous cath NEC             25
7      50–79    Entral infus nutrit sub             22
8      50–79    Packed cell transfusion             13
9        >80           Venous cath NEC             20
10       >80    Packed cell transfusion             13
11       >80  Insert endotracheal tube              8
```

2d. From the above table, it is clear that “Venous cath NEC” is the most common procedure across all ages. The table also indicates that procedure frequency increases with age, with the majority of procedures falling in the 50-79 and >80 age groups. Younger patients (0-19 and 20-49) had fewer procedures, which makes sense because older patients likely require a greater number of surgeries.

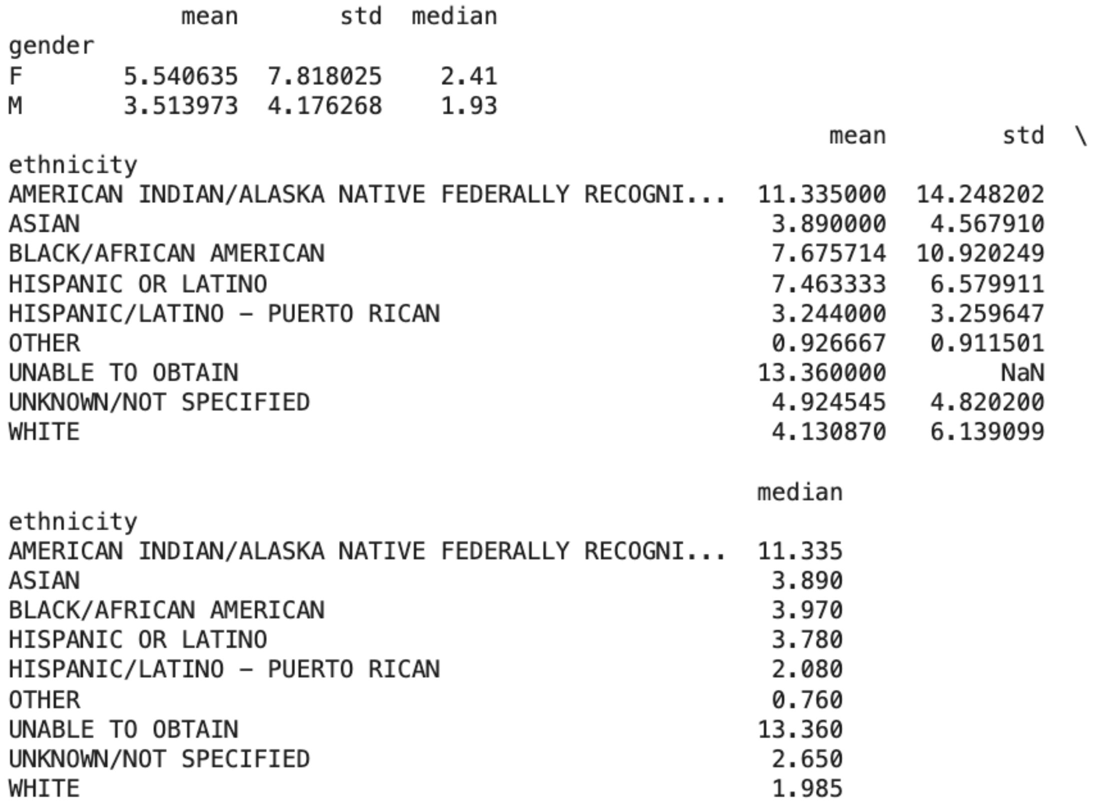
3b. The query joins the ICUSTAYS and ADMISSIONS tables on hadm_id to access the ethnicity data and length of stay. Then, the ADMISSIONS and PATIENTS tables are joined on subject_id to access the gender of the patients. The subject_id, ethnicity, gender, and length of stay (rounded to 2 decimal places) is selected. This dataset shows the length of stay for each patient, along with their ethnicity and gender. Using Python, the mean and standard for gender and ethnicity are printed to determine if there is a difference in the ICU length of stay among gender or ethnicity, and then box plots are printed to visualize the distribution.

3c. The following table shows the first few lines of the dataframe from the SQL query, showing the ethnicity, gender, and length of stay correlated with each patient’s subject_id.



	subject_id	ethnicity	gender	icu_length_of_stay
0	10006	BLACK/AFRICAN AMERICAN	F	1.63
1	10011	UNKNOWN/NOT SPECIFIED	F	13.85
2	10013	UNKNOWN/NOT SPECIFIED	F	2.65
3	10017	WHITE	F	2.14
4	10019	WHITE	M	1.29

The following shows the output from the Python calculations.

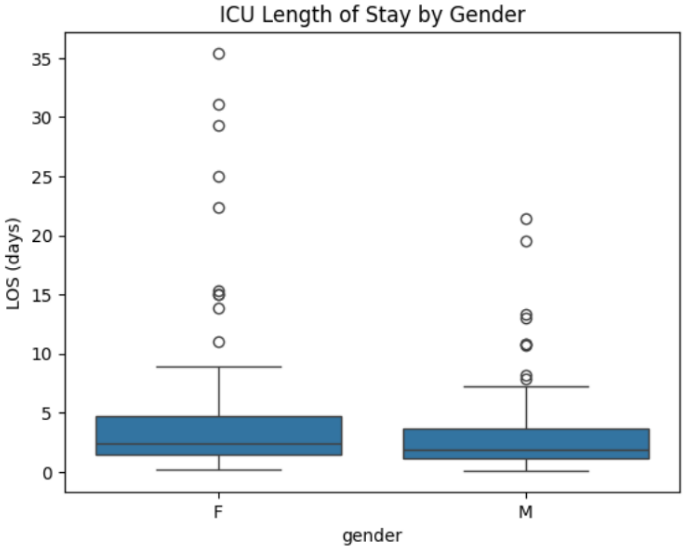


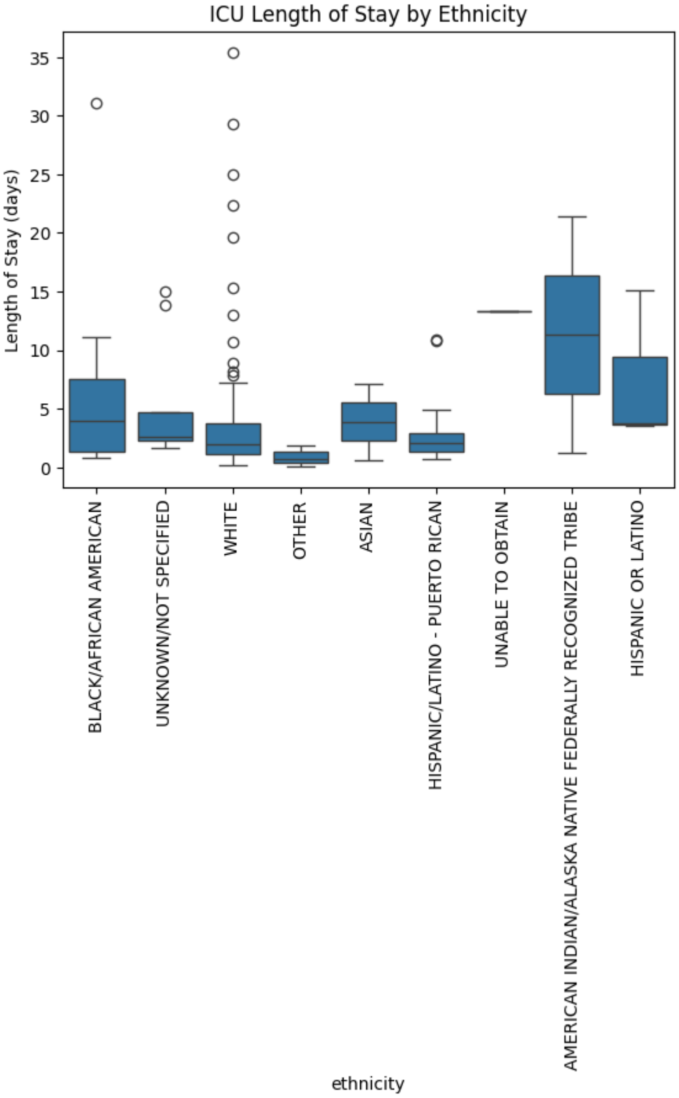
```
            mean       std  median
gender
F       5.540635  7.818025    2.41
M       3.513973  4.176268    1.93
                                                         mean        std  \
ethnicity
AMERICAN INDIAN/ALASKA NATIVE FEDERALLY RECOGNI...  11.335000  14.248202
ASIAN                                                3.890000   4.567910
BLACK/AFRICAN AMERICAN                               7.675714  10.920249
HISPANIC OR LATINO                                   7.463333   6.579911
HISPANIC/LATINO - PUERTO RICAN                       3.244000   3.259647
OTHER                                                0.926667   0.911501
UNABLE TO OBTAIN                                    13.360000        NaN
UNKNOWN/NOT SPECIFIED                                4.924545   4.820200
WHITE                                                4.130870   6.139099

                                                    median
ethnicity
AMERICAN INDIAN/ALASKA NATIVE FEDERALLY RECOGNI...  11.335
ASIAN                                                3.890
BLACK/AFRICAN AMERICAN                               3.970
HISPANIC OR LATINO                                   3.780
HISPANIC/LATINO - PUERTO RICAN                       2.080
OTHER                                                0.760
UNABLE TO OBTAIN                                    13.360
UNKNOWN/NOT SPECIFIED                                2.650
WHITE                                                1.985
```

3d. Based on the calculations comparing the ICU length of stay and gender, it appears that females typically had a longer average ICU stay with greater variability compared to males. This is also shown in the boxplot below, where there are far more outliers with females. The median is still higher for females, although outliers affected the mean calculations by a relatively large amount.

ICU length of stay varied a lot across ethnicities. The highest average was for the Unable to obtain data point, but since there is not standard deviation for this ethnicity and based on the box plot, there is only one data point for this group, indicating that it is not representative and would likely be removed in any further analysis. Without that data point, the American Indian/Alaska Native Federally Recognized Tribe had a much larger mean and median than the other ethnicities, as compared to the groups with the shortest stays of 0.926667 (Other) and 3.244000 (Hispanic/Latino-Puerto Rican). The boxplot shows that there is a lot of outliers among white patients and larger variability in certain ethnicities. Plots for gender and ethnicity are shown below.


ICU Length of Stay by Gender
LOS (days)
0
5
10
15
20
25
30
35
F
M
gender


ICU Length of Stay by Ethnicity
Length of Stay (days)
0
5
10
15
20
25
30
35
BLACK/AFRICAN AMERICAN
UNKNOWN/NOT SPECIFIED
WHITE
OTHER
ASIAN
HISPANIC/LATINO - PUERTO RICAN
UNABLE TO OBTAIN
AMERICAN INDIAN/ALASKA NATIVE FEDERALLY RECOGNIZED TRIBE
HISPANIC OR LATINO
ethnicity

Part II

I, Katie Mowry, acknowledge that no copies of the AWS crendentials file is stored on any publicly accessible location, nor is the file in any way shared with anyone outside of DATA_ENG 300 (Spring 2025).

All code for designing Cassandra tables, uploading data, and queries are shown in the DATA ENG 300 Homework 2 Part II.ipynb file in github. All extraction produced the desired data, matching the results shown above in Part I.

Generative AI Statement

I used Generative AI for help with connecting to my EC2 instance and launching the Docker container. I don’t have the exact prompts saved, but I mainly pasted in the error messages and asked for clarification or potential fixes.