Stats101 Assignment Report

Juejing Han
jhan446@gatech.edu

1 STATS101 ASSIGNMENT REPORT

1.1 Data Exploration (Step 2)

Dataset: Deaths in Custody

Number of Observations: 8373 (8372 after data cleaning)

Number of Variables: 17

Regulated Domain in Law: Housing & Public Accommodation, and it is more closely aligned with Public Accommodation (Civil Rights Act of 1964)

Number of Protected Class Variables: 3

Variable Name	Protected Class	Law
Age	Age	Age Discrimination in Employment Act of 1967 (Over 40)
Gender	Sex	Equal Pay Act of 1963; Civil Rights Act of 1964, 1991
Race	Race	Civil Rights Act of 1964, 1991

Data Cleaning: One entry in the variable "Age" is marked as an unknown value. After removing the entire observation associated with this entry, the dataset consists of **8372** valid observations.

1.2 Relationship between Independent & Dependent Variables (Step 3)

In this step, three independent variables (protected class variables, i.e., Age, Gender, and Race) and two dependent variables (Manner of Death and Custody Status) are examined.

Age is divided into two categories: "under 40" (age < 40) and "40 & Above" (age ≥ 40).

Race has four categories: White, Black, Hispanic, and Other. The "Other" category includes Chinese, Korean, Japanese, Laotian, Filipino, Vietnamese, Cambodian, Asian Indian, Other Asian, Samoan, Hawaiian, American Indian, Pacific Islander, Guamanian, and Other.

Manner of Death is grouped into six categories: Suicide, Natural, Homicide, Accidental, Cannot be Determined (CD), and Other. The "Homicide" category includes Homicide Willful

(Other Inmate), Homicide Justified (Law Enforcement Staff), Homicide Willful (Law Enforcement Staff), and Homicide Justified (Other Inmate). The "Other" category includes Execution, Pending Investigation, and Other.

Custody Status is classified into five categories: Sentenced, Awaiting Booking (AB), Booked-Awaiting Trial (B-AT), Booked-No Charges Filed (B-NCF), and Other. The "Other" category includes In Transit, Process of Arrest, Out to Court, and Other.

Table 1 — Frequency of Age vs Manners of Death (CD: Cannot be Determined).

Age	Homicide	Suicide	Natural	Accident	CD	Other
Under 40	994	482	442	347	59	148
40 & Above	510	374	4493	336	56	131

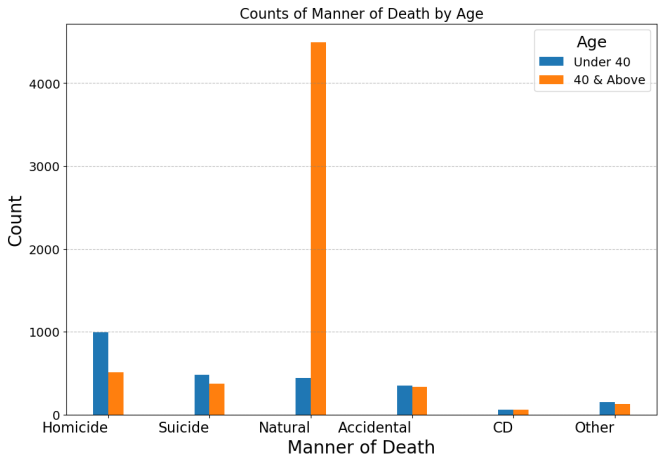

Figure 1 — Frequency of Manner of Death by Age. (CD: Cannot be Determined)

Table 2 — Frequency of Race vs Manners of Death (CD: Cannot be Determined).

Race	Homicide	Suicide	Natural	Accident	CD	Other
White	468	419	2142	265	52	94
Black	294	94	1348	142	29	65
Hispanic	645	263	1211	247	26	101
Other	97	80	234	29	8	19

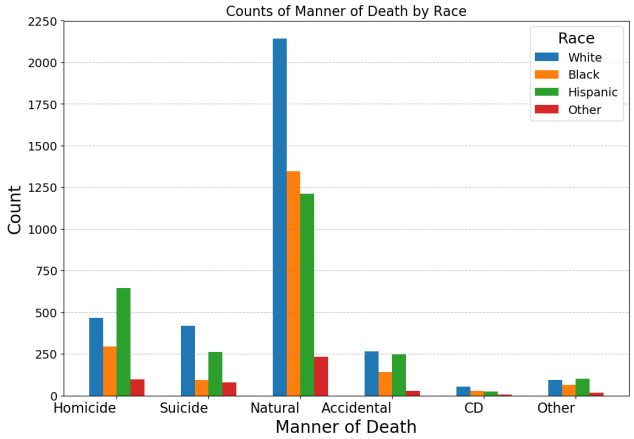

Figure 2 — Frequency of Manner of Death by Race. (CD: Cannot be Determined)

Table 3 — Frequency of Gender vs Manners of Death (CD: Cannot be Determined).

Gender	Homicide	Suicide	Natural	Accident	CD	Other
Male	1452	798	4654	628	106	263
Female	52	58	281	55	9	16

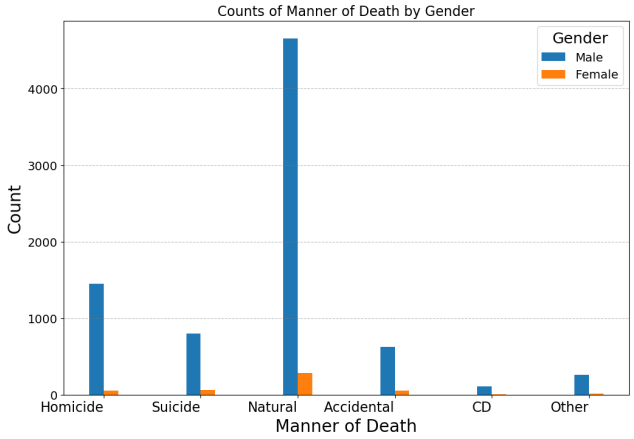

Figure 3 — Frequency of Manner of Death by Gender. (CD: Cannot be Determined)

Table 4 — Frequency of Age vs Custody Status (AB: Awaiting Booking, B-AT: Booked-Waiting Trial, B-NCF: Booked-No Charges Filed).

Age	Sentenced	AB	B-AT	B-NCF	Other
Under 40	777	41	364	78	1212
40 & Above	4418	56	667	125	634

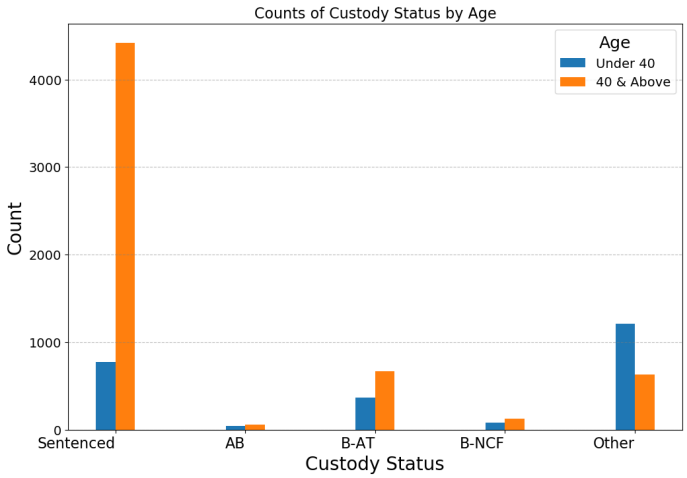

Figure 4 — Frequency of Custody Status by Age (AB: Awaiting Booking, B-AT: Booked-Waiting Trial, B-NCF: Booked-No Charges Filed).

Table 5 — Frequency of Race vs Custody Status (AB: Awaiting Booking, B-AT: Booked-Waiting Trial, B-NCF: Booked-No Charges Filed).

Race	Sentenced	AB	B-AT	B-NCF	Other
White	2249	45	471	92	583
Black	1316	20	211	43	382
Hispanic	1365	31	286	54	757
Other	265	1	63	14	124

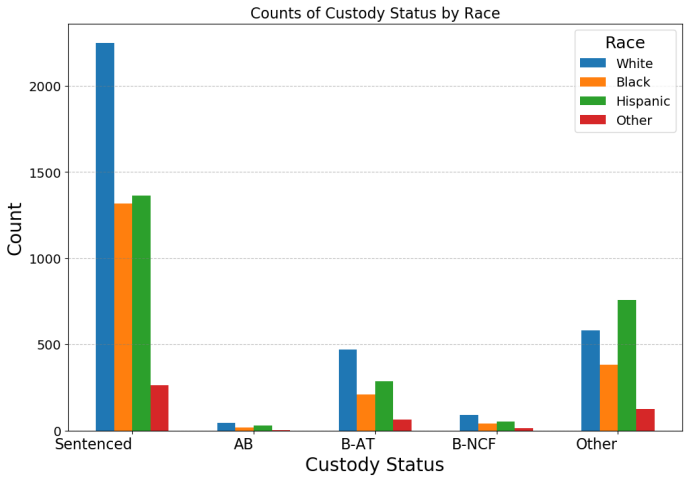

Figure 5 — Frequency of Custody Status by Race (AB: Awaiting Booking, B-AT: Booked-Waiting Trial, B-NCF: Booked-No Charges Filed)

Table 6 — Frequency of Gender vs Custody Status (AB: Awaiting Booking, B-AT: Booked-Waiting Trial, B-NCF: Booked-No Charges Filed).

Gender	Sentenced	AB	B-AT	B-NCF	Other
Male	4961	82	925	171	1762
Female	234	15	106	32	84

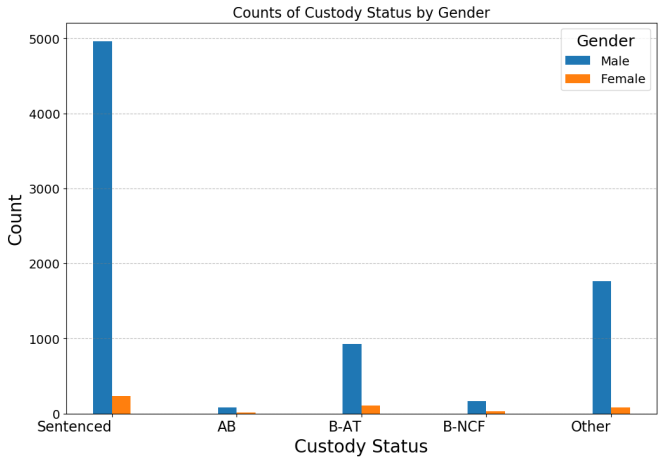

Figure 6 — Frequency of Custody Status by Gender (AB: Awaiting Booking, B-AT: Booked-Waiting Trial, B-NCF: Booked-No Charges Filed).

1.3 Data Manipulation (Step 4)

1) Fair Hypothesis: As shown in Fig. 7, Sentenced Death is not dependent on age. [Manipulations: Used line graph; Increased Scale to +-100; Mapped the ratio of Sentenced Death (i.e., 777/5195 versus 4418/5195); No label on the Y-axis].

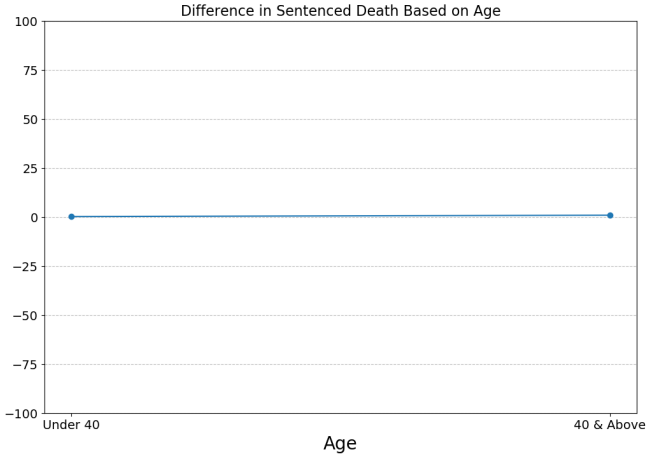

Figure 7 — Difference in Sentenced Death Based on Age.

2) Bias Hypothesis: As illustrated in Fig. 8, there is a significant dependency of Sentenced Death on age. [This hypothesis found straightforward support from the data so did not require much manipulation: Used stacked bar graph; Reduced Scale, which fits the data range; Reworded labels, which represent the corresponding statistics].

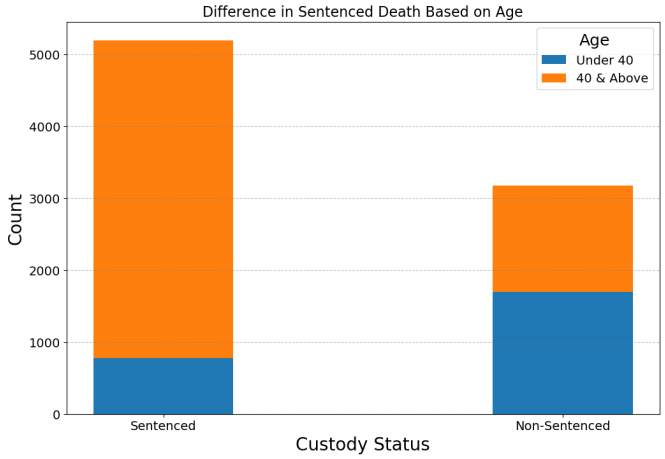

Figure 8 — Difference in Sentenced Death Based on Age.

1.4 Random Sampling (Step 5)

In Section 1.3, Age is chosen as the protected class variable. The numerical values for two Age categories, "Under 40" (2472 entries) and "40 & Above" (5900 entries), are set to 1 and 2, respectively. A comparison of statistics between the original Age group and the reduced Age group (50% reduction with random sampling) is presented in Table 7. Table 7 reveals **a slight-to-no difference in the mean** with **no difference observed in the median or mode**. This observation holds true with different random seeds. The rationale lies in the random sampling method, which offers each entry an equal chance for selection, resulting in the reduced dataset likely sharing a similar distribution with the original data. Besides, given its disproportionate size in the original data, "40 & Above" category is prone to dominate the reduced Age group.

Table 7 — Original Data vs Reduced Data (Statistics of Age Group).

Age	Mean	Median	Mode
Original Data Set	1.70	2.0	2.0
Reduced Data Set	1.71	2.0	2.0
Difference	Slight-to-No Difference	No Difference	No Difference

1.5 Independent & Dependent Variables of Reduced Data (Step 6)

Utilize the reduced data obtained in Section 1.4 to analyze the relationship between the independent variable (Age) and the dependent variable (Custody Status) selected in Section 1.3. While Section 1.3 focused on the specific "Sentenced" category within Custody Status group as the dependent variable, this section examines the entire Custody Status group to provide a comprehensive understanding of the potential impact of random sample selection.

As shown in Table 8, **each pair of independent and dependent variables exhibits a reduction in counts**, approximately half the size compared to the statistics of the original data (presented in parentheses in Table 8 for reference). This reduction aligns with the 50% reduced sampling.

Table 8 — Frequency of Age vs Custody Status of Reduced Data (AB: Awaiting Booking, B-AT: Booked-Waiting Trial, B-NCF: Booked-No Charges Filed). Original frequency is listed in parentheses for reference.

Age	Sentenced	AB	B-AT	B-NCF	Other
Under 40	389 (777)	21 (41)	187 (364)	40 (78)	596 (1212)
40 & Above	2196 (4418)	28 (56)	347 (667)	58 (125)	324 (634)

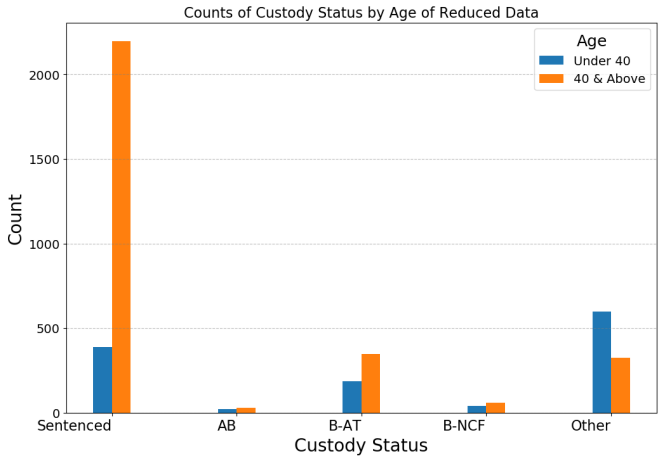

Figure 9 — Frequency of Custody Status by Age of Reduced Data (AB: Awaiting Booking, B-AT: Booked-Waiting Trial, B-NCF: Booked-No Charges Filed).

Fig. 9 further demonstrates that **despite the reduced frequencies in each pair of independent and dependent variables within the reduced data compared to the original data (Fig. 4), there are no discernible differences between the original and reduced datasets in terms of the pattern and distribution of each pair of variables**. This consistency holds true across different random seeds. The reason remains the same as stated in Section 1.4 – the random sampling

method ensures each sample has an equal opportunity for selection, so the reduced data shares a similar distribution with the original dataset.

From the comparison above, it is evident that after utilizing the random sampling process, **members associated with the protected class variable would NOT benefit or be harmed**. **This is because** random sampling gives each member an equal chance to be selected for inclusion in the sample. Random sampling does not favor or disfavor any specific group. It contributes to ensuring that the sample reflects the characteristics of the overall population.