PSTAT100 Mid Quarter Project

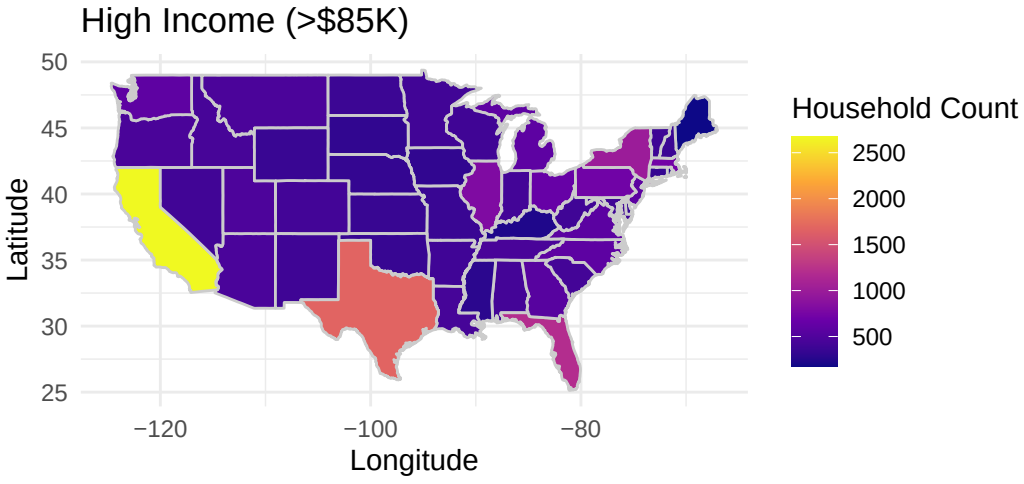
Austin Zhang (xinhaozhang@umail.ucsb.edu)
Ruiqi Li (ruiqi_li@ucsb.edu)
Chris Orellana (c_orellana@ucsb.edu)
Andrew Guerra (andrewguerra@umail.ucsb.edu)

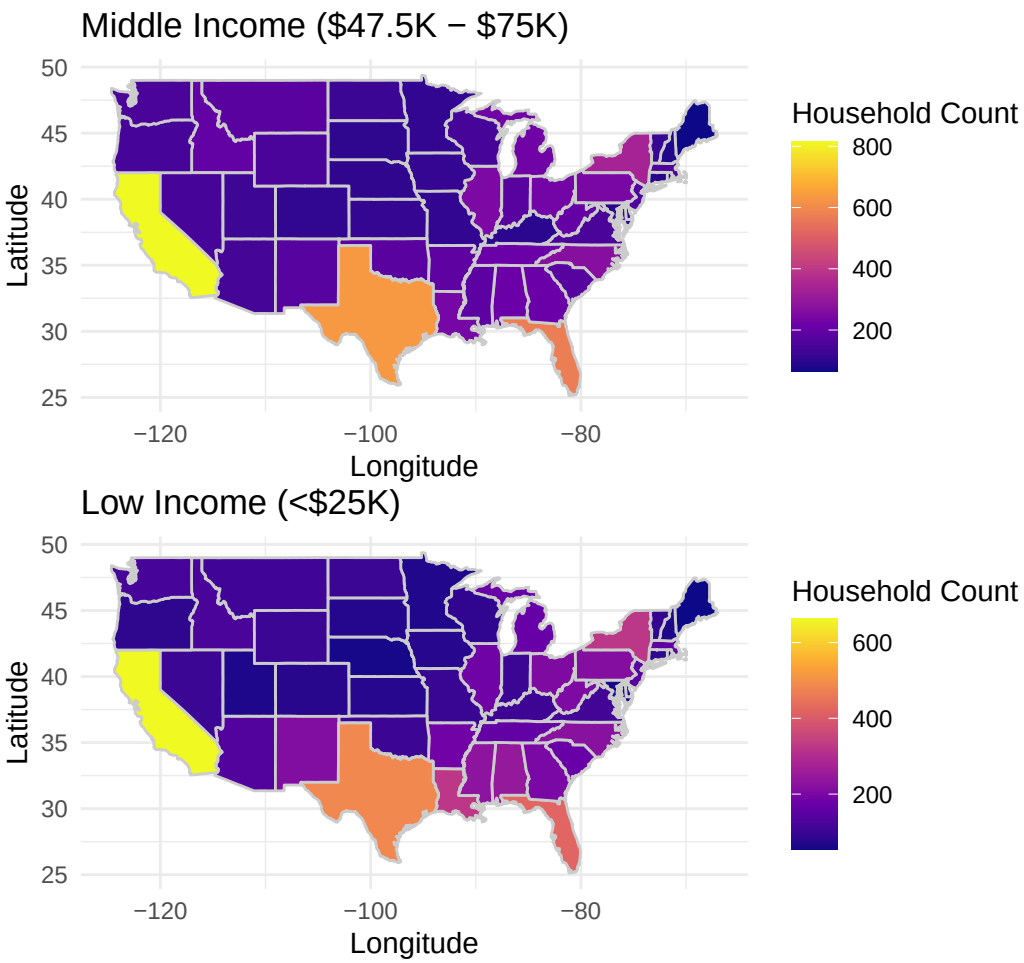
2025-07-12

! Part 1: Household Level

Before getting started, we need to briefly introduce Core Based Statistical Area (CBSA). According to U.S. Census Bureau, it is a geographic region that contains one or more counties anchored by an urban center of at least **10,000** people. There are 2 types of CBSA: metropolitan statistical areas (**population > 50,000**) & micropolitan statistical areas (**10,000 < population < 49,999**). Usually, CBSA codes are used to compare and identify geographic regions within the US. We'll see how they can be applied to our research very soon.

Household Income distribution Analysis





To analyze household income distribution across the US, we divided households into three income levels: low income (< $25,000), middle income (approximately $47,500–$75,000), and high income (>= $85,000). The maps above show the number of households in each state for every income group.

Let's start with the overall trend. In general, states with larger populations—such as California, Texas, and Florida—tend to have higher household counts across all income levels.

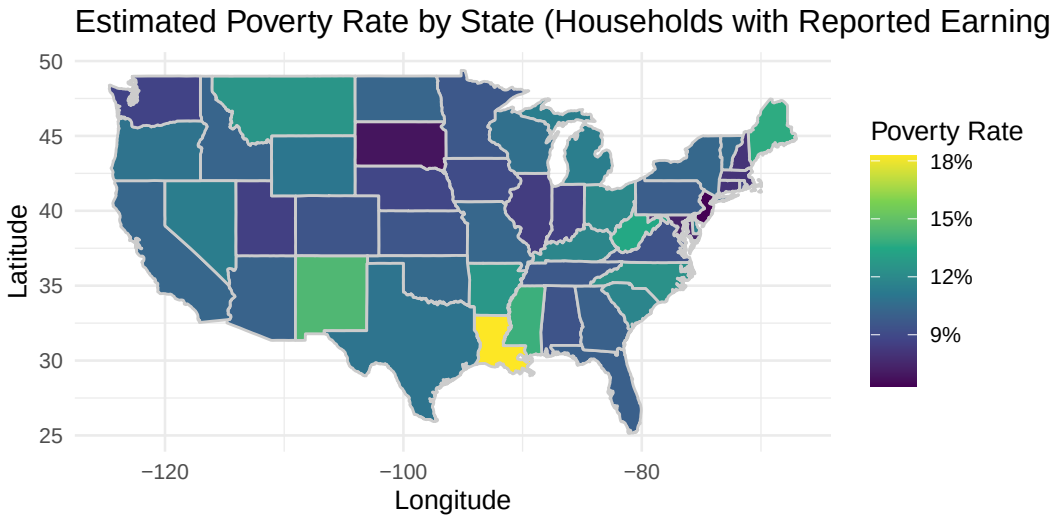
Then, we can take a deeper look at each map:

- **High-income households** are more prevalent in states with larger urban economies and higher costs of living, including California, New York, and Washington.
- **Middle-income households** seem to evenly distributed nationwide, likely representing the income range of most American households.

- **Low-income households** mainly concentrates in Southern states like Mississippi, Louisiana, and Alabama, which reflects broader patterns of regional economic disparity.

All in all, these maps highlight both the role of population size in shaping household counts and the geographic variation in income distribution across the country.
Now, let's take a closer look at low-income households. Specifically, what is the proportion of households within each state that fall below poverty lines?

Households Below the Poverty Line

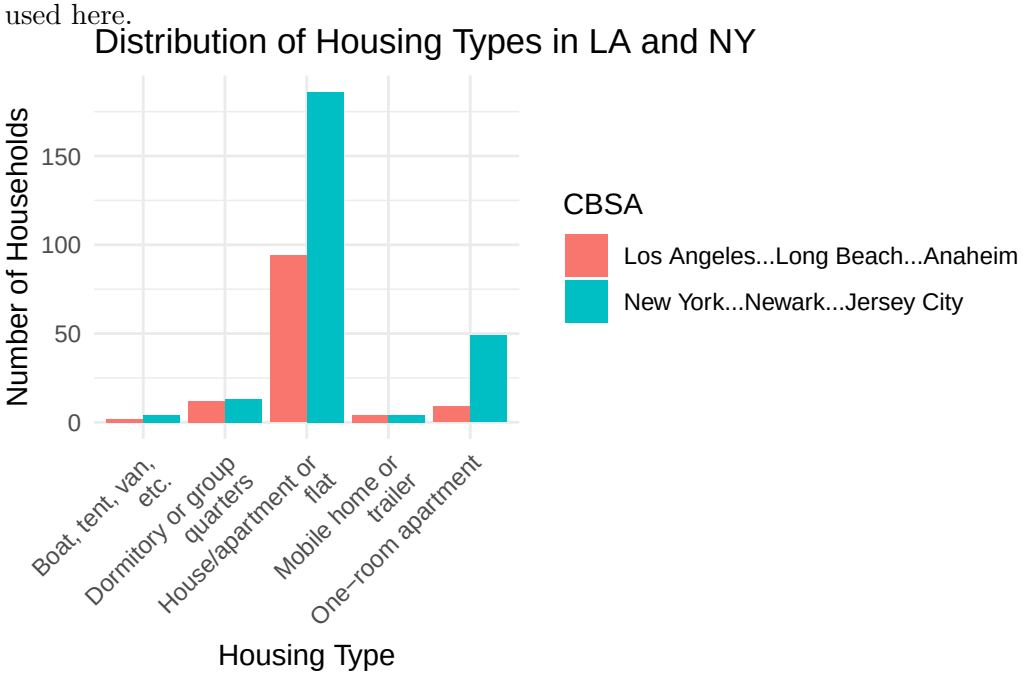


Based on HEARNVAL and 2025 HHS Poverty Guidelines

To find out the households that fall below the poverty line, we took a look at the given 2025 U.S. Health and Human Services (HHS) poverty guidelines. A function poverty_threshold() is created to define and assign the appropriate poverty threshold based on household size. In the case of households that are larger than 8 people, we extended the guideline by adding $5,000 for each additional person to compensate for the information not given.
In the given data dictionary, the income variable HEARNVAL reports total household earnings. Although these earnings don't capture all sources of household income, such as benefits or financial assistance, we treated the HEARNVAL variable as the best and most appropriate approximation of household income for our poverty analysis. Due to the fact that HEARNVAL reports the exact earnings instead of the income brackets, we compared it directly against the calculated poverty threshold for each household size.
Next, we restricted our data analysis to the households that reported any earnings at all by then filtering the households where the universe variables **HINC_WS, HINC_SE, or HINC_FR** indicated positive income. This exclusion avoids households with no earnings, thereby skewing the poverty rate estimation, as our variable analysis focuses solely on reported earnings.

Finally, we aggregated households at the state level to calculate the proportion of households falling below the poverty threshold within each state. The result is visualized by using a diverging color palette map, where each state's fill represents its estimated poverty rate.

Next, we'll go over a more specific case: the proportion of single individuals living in houses and/or apartments that fall below the poverty line in the **Los Angeles-Long Beach-Anaheim and the New York-Newark-Jersey City areas.** CBSA will be used here.



The bar chart indicates the housing types from Los Angeles–Long Beach–Anaheim and New York–Newark–Jersey City CBSAs. Here are what we have observed:

- The majority of households in both regions reside in **homes** or **apartments**. The number of households in New York is significantly higher.
- **Los Angeles** exhibits more **trailers or mobile homes**, probably as a result of its more dispersed terrain.
- Because of its increased population density and constrained space, **New York** has more **one-room flats** and **dormitory-style housing**.
- **Few households** in any area reside in **non-traditional dwellings, such as tents, vans, or boats**. Each city's unique housing structure and usage are reflected in these patterns.

Table 1: Poverty Rate in LA and NY

Region	Poverty_Rate
Los Angeles	9.216590
New York	8.626198

Now, let's turn to the calculation for proportion of single individuals living in houses and/or apartments that fall below the poverty line. Here, we used the hhpub24.csv dataset and filtered for: CBSA area codes(LA: GTCBSA == 31080, NY: GTCBSA == 35620), Single-person households(filtered with H_NUMPER == 1), Valid income values(Filtered out invalid or missing entries using HEARNVAL > 0), Poverty threshold(defined a poverty threshold of $15,000 annual income), and poverty rate calculation. Los Angeles (**9.22%**) and New York (**8.63%**) have relatively equal poverty rates for single-person households, indicating that living alone in both cities poses comparable financial difficulties. This phenomenon is caused by the high cost of living in both areas. The slight difference in poverty rates may result from differences in regional support networks, such as public benefits or housing help, or the differences in housing availability or employment markets.

Any Missing Values?

Table 2: Summary of Missing Values

Dataset	Variable	Missing_Count	Missing_Proportion
household_income	state_name	768	0.0179721
household_income	region	768	0.0179721

Lastly, we need to check the missing values in the dataset. We create a function named missing_check to find variables that contain missing values and calculate the proportion of these missing values. According to the table offered above, state_name and region in dataset household_income are the one that contains missing values with proportion of approximately 1.797%. However, no imputation has occured because we are only using filtering and joins.

! Part 2: Family Level

Having explored patterns and disparities at the household level, we now shift our focus to the familial level. More specifically, we aim to investigate the relationship between medical-related expenditures and poverty guidelines among same-sex couples.
First, we begin by determining the distribution of families with respect to the poverty line. The variable "FAMLIS" denotes the ratio of family income to poverty threshold.

Using this metric, families included in the poverty status determination are categorized from 1 to 4, indicating the following:

- **FAMLIS = 1:** denotes families living below poverty level.
- **FAMLIS = 2:** denotes families living from 100% to 124% above the poverty level.
- **FAMLIS = 3:** denotes families living 125% to 149% above the poverty level.
- **FAMLIS = 4:** denotes families living 150% and above the poverty level

Using these values, we created a table to determine the distribution of same-sex couples among FAMLIS values.
Based on our table, we observe that a total of 343 same-sex couple families were included in the sample. Of these 343 families:

- **325 (94.75%)** live at or above 150% of the poverty level, suggesting that the vast majority of same-sex couple families in the sample are economically stable.
- **10 same-sex couple families (2.92%)** live below the poverty level, followed by **5 families (1.46%)** living at or to about **124%** of the poverty level.
- **3 same-sex couple families (0.87%)** live from 125% to 149% of the poverty level.

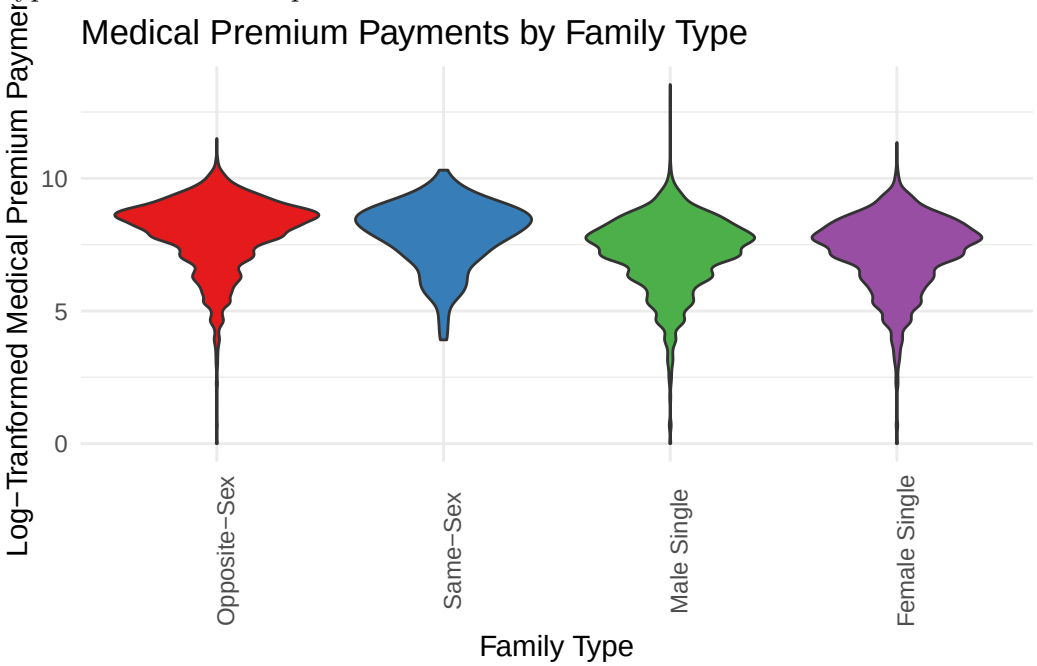
Table 3: Poverty Ratio Among Same-Sex Couple Families

FAMLIS	n	distribution
1	10	2.92
2	5	1.46
3	3	0.87
4	325	94.75

Since 10 same-sex couple families of the 343 sampled live below the poverty line, we can infer that 333 families **(97.08%)** that were sampled live above the poverty line.

Medical Premiums Paid Among Different Family Types

To investigate differences in the amount different family types paid in medical premiums, we decide to plot our data using side-by-side violin plots. We applied a log-transformation on the amount paid to normalize the inherent right-skewness and grouped each family type to distinct violin plots.



According to the graph, most of the distribution is between 5 to 10 on the log-scale across all family types. While no family type shows consitent higher or lower premium payments, there are some slight variabilities in the distribution and range across all family types. Same-sex couple families show the smallest variation of premium payments, ranging from about 4 to 10 on the log-scale (**approximately $54.60–$22,026.47**). Conversely, the family type with the greatest variation are single male families, ranging from 0 to about 17 on the log scale (**approxiately $0–$24,154,952.75**). single female families closely follow this variation, ranging from 0 to about 13 on the log scale (**approximately $0–$442,413.39**). Interestingly, opposite-sex couples also range from 0 to about 13 on the log scale (**approximately $0–$442,413.39**), though the distribution among opposite-sex couples suggests their premium payments are higher, on average. Though the range varies widely, a majority of payments are within 5 to 10 on the log-scale (**approximately $148.41–$22,026.47**).

Imputed Values

To determine the proportion of values that were imputed, we constructed a table based on the variable "I_FHIPVAL", which assigns a number 0 to 3 indicating the following:

- **I_FHIPVAL = 0:** denotes values that were not imputed.
- **I_FHIPVAL = 1:** denotes values from Hotdeck Imputation.
- **I_FHIPVAL = 2:** denotes values from Logical Imputation.
- **I_FHIPVAL = 3:** denotes values from Whole Unit Imputation.

Based on these values, we computed the proportion to find exactly how many values were imputed.
According to the table below:

- 37,668 (**57.74%**) of the total amount each family paid in premiums was not imputed. Thus, 27,569 **(42.26%)** of values were imputed.
- Of the imputed premiums paid, 11,591 payments (**17.77%**) were imputed by a Hotdeck Imputation.
- Another 447 payments (**0.69%)** were imputed by a Logical Imputation.
- Whole Unit Imputation appeared to be the most used imputation method, with 15,531 payments (**23.81%**) resulting from it.

Table 4: Proportion of Payments by I_FHIPVAL

I_FHIPVAL	n	proportion
0	37668	57.74
1	11591	17.77
2	447	0.69
3	15531	23.81

To assess the distribution of imputation methods, we created another table assuming imputation methods used:

Table 5: Distribution of Payments Imputed

I_FHIPVAL	n	proportion
1	11591	42.04
2	447	1.62
3	15531	56.34

Among payments that were imputed, we observe that a majority of values imputed were done so with the Whole Deck Unit method, making up about **56.34%** of all imputed values. Following this method is Hotdeck Imputation, consisting of about **42.04%** of all imputed values. Lastly, the least used method is Logical Imputation, comprising of only about **1.62%** of all imputed values.

Hotdeck Imputation
Using Google, we found that Hotdeck Imputation is a type of technique that assigns a value by inference based on the values of other, similar data points that are found within a given dataset. It handles missing data by replacing missing values with values from other similar data points and prioritizes preserving the current distribution of the data while being able to maintain relations across multiple variables (Andridge & Little).
While Hotdeck Imputation is generally effective, it leaves room for data to be affected more critically by bias. If sampling data points that help reassign missing values was not done properly, we risk over-representing bias in our results (Andridge & Little).

! References

Andridge, Rebecca R, and Roderick J A Little. "A Review of Hot Deck Imputation for Survey Non-response." *International statistical review = Revue internationale de statistique* vol. 78,1 (2010): 40-64. doi:10.1111/j.1751-5823.2010.00103.x

Appendices

Appendix A: Part I

```
library(tidyverse)
states <- map_data("state")

ggplot() +
  geom_polygon(data = states,
               aes(x = long, y = lat, group = group),
               fill = "grey90",
               colour = "grey50") +
  coord_quickmap() +
  theme_minimal()

# Question 2 (HHINC Variable used Total household income)

library(tidyverse)
library(maps)

household <- read.csv("data/hhpub24.csv")
fips_abbrev <- read.csv("data/fips_abbrev.csv", header = FALSE, skip = 1,
                        col.names = c("State", "FIPS", "USPS"))

fips_abbrev <- fips_abbrev %>%
  mutate(
    fips = as.integer(FIPS),
    abbrev = toupper(trimws(as.character(USPS)))
  ) %>%
  select(fips, abbrev)

# Gathering info about the states
state_lookup <- tibble(
  abbrev = toupper(state.abb),
  state_name = state.name,
  region = tolower(state.name)
)
```

```r
# Merge and assign income group
household <- household %>%
  mutate(GESTFIPS = as.integer(GESTFIPS)) %>%
  left_join(fips_abbrev, by = c("GESTFIPS" = "fips")) %>%
  left_join(state_lookup, by = "abbrev") %>%
  mutate(
    income_group = case_when(
      HHINC >= 1 & HHINC <= 10 ~ "Low Income (< $25K)",
      HHINC >= 20 & HHINC <= 30 ~ "Middle Income ($47.5K-$75K)",
      HHINC >= 35 & HHINC <= 41 ~ "High Income (  $85K)",
      TRUE ~ NA_character_
    )
  )

# Summarize household counts
income_counts <- household %>%
  filter(!is.na(income_group), !is.na(region)) %>%
  group_by(region, income_group) %>%
  summarise(count = n(), .groups = "drop")

# Merge with map data
states <- map_data("state")
income_map_data <- states %>%
  left_join(income_counts, by = "region")

# Plot the map
high_income <- income_map_data %>%
  filter(income_group == "High Income (  $85K)")
middle_income <- income_map_data %>%
  filter(income_group == "Middle Income ($47.5K-$75K)")
low_income <- income_map_data %>%
  filter(income_group == "Low Income (< $25K)")

# Plot 1: High Income
ggplot(high_income, aes(x = long, y = lat, group = group, fill = count)) +
  geom_polygon(color = "gray80") +
  coord_quickmap() +
  scale_fill_viridis_c(option = "plasma", na.value = "gray90") +
  labs(title = "High Income (>$85K)",
       fill = "Household Count",
       x = "Longitude",
       y = "Latitude") +
  theme_minimal()
```

```r
# Plot 2: Middle Income
ggplot(middle_income, aes(x = long, y = lat, group = group, fill = count)) +
  geom_polygon(color = "gray80") +
  coord_quickmap() +
  scale_fill_viridis_c(option = "plasma", na.value = "gray90") +
  labs(title = "Middle Income ($47.5K - $75K)",
       fill = "Household Count",
       x = "Longitude",
       y = "Latitude") +
  theme_minimal()

# Plot 3: Low Income
ggplot(low_income, aes(x = long, y = lat, group = group, fill = count)) +
  geom_polygon(color = "gray80") +
  coord_quickmap() +
  scale_fill_viridis_c(option = "plasma", na.value = "gray90") +
  labs(title = "Low Income (<$25K)",
       fill = "Household Count",
       x = "Longitude",
       y = "Latitude") +
  theme_minimal()

# Question 3 (HEARNVAL Variable, total household earnings)
household <- read.csv("data/hhpub24.csv")
fips_abbrev <- read.csv("data/fips_abbrev.csv", header = FALSE, skip = 1,
                        col.names = c("State", "FIPS", "USPS"))

poverty_threshold <- function(hh_size) {
  case_when(
    hh_size == 1 ~ 15650,
    hh_size == 2 ~ 21150,
    hh_size == 3 ~ 26650,
    hh_size == 4 ~ 32150,
    hh_size == 5 ~ 37650,
    hh_size == 6 ~ 43150,
    hh_size == 7 ~ 48650,
    hh_size == 8 ~ 54150,
    hh_size > 8 ~ 54150 + 5000 * (hh_size - 8),
    TRUE ~ NA_real_
  )
}
```

```
# Filter to only households that reported earnings
household_income <- household %>%
  filter(HINC_WS == 1 | HINC_SE == 1 | HINC_FR == 1)

# Compute poverty line and flag below-poverty households
household_income <- household_income %>%
  mutate(
    estimated_income = HEARNVAL,
    hh_size = H_NUMPER,
    poverty_line = poverty_threshold(hh_size),
    below_poverty = ifelse(estimated_income < poverty_line, 1, 0)
  )

# Clean fips_abbrev
fips_abbrev <- fips_abbrev %>%
  mutate(
    fips = as.integer(FIPS),
    abbrev = toupper(trimws(as.character(USPS)))
  ) %>%
  select(fips, abbrev)

# Create state name lookup table
state_lookup <- tibble(
  abbrev = toupper(state.abb),
  state_name = state.name,
  region = tolower(state.name)
)

# Add region/state info to household_income
household_income <- household_income %>%
  left_join(fips_abbrev, by = c("GESTFIPS" = "fips")) %>%
  left_join(state_lookup, by = "abbrev")

# Summarize by state
poverty_summary <- household_income %>%
  filter(!is.na(region)) %>%
  group_by(region) %>%
  summarise(
    total_households = n(),
    poor_households = sum(below_poverty, na.rm = TRUE),
    poverty_rate = poor_households / total_households
  )
```

```r
# Map
poverty_map_data <- map_data("state") %>%
  left_join(poverty_summary, by = "region")

ggplot(poverty_map_data, aes(x = long, y = lat, group = group,
                             fill = poverty_rate)) +
  geom_polygon(color = "gray80") +
  scale_fill_viridis_c(name = "Poverty Rate",
                       labels = scales::percent,
                       na.value = "gray90") +
  coord_quickmap() +
  labs(
    title = "Estimated Poverty Rate by State (Households with Reported Earnings)",
    caption = "Based on HEARNVAL and 2025 HHS Poverty Guidelines",
    x = "Longitude",
    y = "Latitude"
  ) +
  theme_minimal(base_size = 10)

# Question 4
cbsa_codes <- c(31080, 35620)

housing_labels <- data.frame(
  H_TYPEBC = c(1, 2, 3, 4, 5),
  housing_type = c(
    "House/apartment or flat",
    "Mobile home or trailer",
    "One-room apartment",
    "Boat, tent, van, etc.",
    "Dormitory or group quarters"
  )
)

housing_cbsa <- household %>%
  filter(GTCBSA %in% cbsa_codes, H_TYPEBC %in% 1:5) %>%
  left_join(housing_labels, by = "H_TYPEBC") %>%
  mutate(cbsa_name = case_when(
    GTCBSA == 31080 ~ "Los Angeles-Long Beach-Anaheim",
    GTCBSA == 35620 ~ "New York-Newark-Jersey City"
  ))
```

```r
ggplot(housing_cbsa, aes(x = housing_type, fill = cbsa_name)) +
  geom_bar(position = "dodge") +
  labs(
    title = "Distribution of Housing Types in LA and NY",
    x = "Housing Type",
    y = "Number of Households",
    fill = "CBSA"
  ) +
  scale_x_discrete(labels = function(x) str_wrap(x, width = 20)) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))

# Question 5
la_data <- household %>%
  filter(GTCBSA == 31080, H_NUMPER == 1)

poverty_threshold <- 15000

la_poverty_rate <- la_data %>%
  filter(HEARNVAL > 0) %>%
  summarise(
    below_poverty = sum(HEARNVAL < poverty_threshold),
    total = n(),
    rate = below_poverty / total * 100
  )

ny_data <- household %>%
  filter(GTCBSA == 35620, H_NUMPER == 1)

ny_poverty_rate <- ny_data %>%
  filter(HEARNVAL > 0) %>%
  summarise(
    below_poverty = sum(HEARNVAL < poverty_threshold),
    total = n(),
    rate = below_poverty / total * 100
  )

poverty_table <- bind_rows(
  la_poverty_rate %>%
    mutate(Region = "Los Angeles"),
  ny_poverty_rate %>%
    mutate(Region = "New York")
) %>%
  select(Region, Poverty_Rate = rate)
```

```
knitr::kable(poverty_table,
             digits = 2, caption = "Poverty Rate in LA and NY")

# Question 6
library(knitr)
missing_check <- function(df, name = "Dataset") {
  missing_counts <- sapply(df, function(x) sum(is.na(x)))
  missing_props <- sapply(df, function(x) mean(is.na(x)))
  tibble(
    Dataset = name,
    Variable = names(missing_counts),
    Missing_Count = missing_counts,
    Missing_Proportion = missing_props
  ) %>%
    filter(Missing_Count > 0) %>%
    arrange(desc(Missing_Proportion))
}

household_missvalue <- missing_check(household, "household")
income_missvalue <- missing_check(household_income, "household_income")
cbsa_missvalue <- missing_check(housing_cbsa, "housing_cbsa")

missing_summary <- bind_rows(household_missvalue,
                             income_missvalue, cbsa_missvalue)
kable(missing_summary, caption = "Summary of Missing Values")
```

Appendix B: Part II

```
families <- read.csv("data/ffpub24.csv") # Loading Our Data
```

```
# Question 1 : Distribution of same-sex families
families %>%
  filter(FKINDEX == 2) %>% # Selects same-sex couples
  count(FAMLIS) %>%
  mutate(distribution = n / sum(n) * 100) %>%
  knitr::kable(,
    caption = "Poverty Ratio Among Same-Sex Couple Families",
    digits = 2) # Table Format
```

```r
# Question 2 : Violin Plot of Medical Premiums Paid by Family Type
families %>%
  filter(!is.na(FHIP_VAL) & FHIP_VAL > 0) %>%
  mutate(FKINDEX = factor(FKINDEX, # Changing x-axis labels
        levels = c(1, 2, 3, 4),
        labels = c("Opposite-Sex","Same-Sex",
                   "Male Single", "Female Single"))) %>%
  ggplot(aes(x = FKINDEX, y = log(FHIP_VAL),
             group = FKINDEX, fill = FKINDEX)) +
  geom_violin()  +
  scale_fill_brewer(palette = "Set1") +
  labs(
    title = "Medical Premium Payments by Family Type",
    y = "Log-Tranformed Medical Premium Payments",
    x = "Family Type"
  ) +
  theme_minimal() +
  theme(legend.position = "none",
        axis.text.x = element_text(angle = 90)
  )
```

```r
# Question 3 : Proportion of Values Imputed
families %>%
  filter(!is.na(I_FHIPVAL) & I_FHIPVAL != -1) %>%
  count(I_FHIPVAL) %>%
  mutate(proportion = n / sum(n) * 100) %>%
  knitr::kable(,
               caption = "Proportion of Payments by I_FHIPVAL",
               digits = 2)
```

```r
# Question 3 : Distribution of Imputation Methods
families %>%
  filter(!is.na(I_FHIPVAL) & I_FHIPVAL != 0) %>%
  count(I_FHIPVAL) %>%
  mutate(proportion = n / sum(n) * 100) %>%
  knitr::kable(,
               caption = "Distribution of Payments Imputed",
               digits = 2)
```