Garlock520FinalProjectMilestone1and2

Matt Garlock

2024-05-19

Introduction

The Financial Burden of Childcare in Florida

The cost of childcare in Florida has become a pressing issue for many families, impacting their economic stability and overall well-being. Rising childcare expenses force families to make difficult financial decisions, affecting savings, lifestyle, and long-term economic health. This project explores the financial burden of childcare on households in Florida by comparing it to their income levels.

Understanding this dynamic is crucial for policymakers, families, and economic planners. It provides insights into the economic challenges faced by families and helps in formulating policies to alleviate this burden.

Data science is particularly suited to this topic due to the need for comprehensive analysis of large, complex datasets involving economic and demographic variables. By leveraging data science techniques, we can uncover patterns, trends, and correlations that are not immediately apparent, providing a deeper understanding of the issue.

Research Questions

1. **What percentage of median household income is spent on childcare in different regions of Florida?** This question quantifies the financial burden by calculating the proportion of income allocated to childcare expenses across various regions.
2. **How do childcare costs in Florida compare to other states?** By comparing childcare costs in Florida to those in other states, we can assess whether Florida's costs are higher, lower, or on par with the national average.
3. **Are there noticeable trends in childcare costs relative to income changes over the past decade?** This question explores the temporal dynamics of childcare costs and income levels to identify trends and shifts over time.
4. **What factors contribute to variations in childcare costs within Florida?** Identifying factors influencing childcare costs can help in understanding regional variations and targeting interventions.
5. **Can we predict future trends in childcare costs and their impact on family finances?** By using predictive modeling, we can forecast future childcare costs and assess their potential impact on families, aiding in proactive planning and policy formulation.

Approach

The analysis will involve a multi-faceted approach:

- **Descriptive Statistics:** Summarize current data on income and childcare costs.
- **Comparative Analysis:** Compare costs across regions within Florida and with other states.
- **Correlational Studies:** Identify factors linked with high childcare costs.
- **Predictive Modeling:** Forecast future trends using regression models.

Data

Dataset 1: Florida Household Income Data

- Source: U.S. Census Bureau
- Description: Median household income statistics by county.
- Variables: County, Year, Median Income
- Purpose: Understanding income distribution across Florida.

Dataset 2: Childcare Cost Data by State

- Source: Child Care Aware of America
- Description: Annual report on the cost of childcare by state.
- Variables: State, Year, Average Childcare Cost
- Purpose: Comparative analysis of childcare costs across states.

Dataset 3: Demographic and Employment Data

- Source: Bureau of Labor Statistics
- Description: Employment status, number of children, and other demographics.
- Variables: Employment rate, Number of children, Age groups, Educational level
- Purpose: Analyzing factors influencing income levels and childcare needs.

Required Packages

- 'tidyverse' # for data manipulation and visualization, providing a comprehensive suite of tools for data analysis.
- 'lubridate' # for handling dates, essential for working with time series data and temporal analysis.
- `ggplot2` # for creating advanced graphical representations, enabling the creation of detailed and informative visualizations.
- `caret` or `forecast`# for predictive modeling, offering tools for building and evaluating predictive models.

Plots and Tables

- **Income vs. Childcare Costs**: Scatter plots and line graphs to visualize trends over time and compare income levels to childcare costs.
- **Heatmaps**: To show childcare cost variability across different regions, providing a visual representation of regional disparities.
- **Regression Diagnostics**: Plots to assess the fit of predictive models, ensuring the accuracy and reliability of forecasts.

Learning Needs

- Geospacial Analysis:Further understanding of geospatial analysis to better interpret regional data variations and create more detailed maps and visualizations.
- Predictive Analytics: Enhancing skills in predictive analytics, particularly in time series forecasting, to build more accurate and robust predictive models.

Future Steps

- Deepening the analysis to incorporate more nuanced socio-economic variables tp provide a comprehensice analysis of the factors affecting childcare costs..
- Collaborating with local policymakers to discuss findings and implications, and to formulate policies based on the insights gained from the analysis.

- Continuously improve the predictive models by incorporating new data and refining the methodologies used.

```
# Load necessary libraries
library(readxl)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggplot2)

# Import the cleaned datasets
fl_household_demo <- read_excel("/Users/mattgarlock/Downloads/Cleaned_Florida_Household_Demographic.xlsx
non_vital_rates <- read_excel("/Users/mattgarlock/Downloads/Cleaned_NonVitalRateOnlyInd_TenYrsReport.xls
ndcp_estimates <- read_excel("/Users/mattgarlock/Downloads/Cleaned_NDCP_State_Level_Estimates_2018_2023

# Display structure of the datasets
str(fl_household_demo)
```

```
## tibble [6 x 10] (S3: tbl_df/tbl/data.frame)
##  $ Total Population            : num [1:6] 21339762 NA NA NA NA ...
##  $ Total Household             : num [1:6] 8157420 NA NA NA NA ...
##  $ Type of Household           : chr [1:6] "Family" "Non-Family" NA NA ...
##  $ Total of Type of Household  : num [1:6] 5274491 2882929 NA NA NA ...
##  $ Children Households         : chr [1:6] "With" "Without" NA NA ...
##  $ Total of Children Households: num [1:6] 2196679 5960741 NA NA NA ...
##  $ Education Level             : chr [1:6] "No High School" "Some High School" "Some College" "Assoc
##  $ Total Education Level       : num [1:6] 707827 6062899 4543900 1715257 3210403 ...
##  $ Marital Status              : chr [1:6] "Never Married" "Married" "Separated" "Widowed" ...
##  $ Marital Status Total        : num [1:6] 5606403 8747349 358544 1169892 2297142 ...
```

```
str(non_vital_rates)
```

```
## tibble [68 x 11] (S3: tbl_df/tbl/data.frame)
##  $ County: chr [1:68] "Florida" "Alachua" "Baker" "Bay" ...
##  $ 2022  : num [1:68] 67917 57566 67872 65999 54759 ...
##  $ 2021  : num [1:68] 61777 53314 63860 60473 48803 ...
##  $ 2020  : num [1:68] 57703 50089 62299 56483 43580 ...
##  $ 2019  : num [1:68] 55660 49689 63275 54316 45921 ...
##  $ 2018  : num [1:68] 53267 49078 61769 51829 46197 ...
##  $ 2017  : num [1:68] 50883 45478 59506 50283 46106 ...
##  $ 2016  : num [1:68] 48900 44702 53327 48577 43373 ...
##  $ 2015  : num [1:68] 47507 43073 47121 47368 41606 ...
##  $ 2014  : num [1:68] 47212 42045 46865 47274 40481 ...
##  $ 2013  : num [1:68] 46956 42149 49236 47461 40259 ...
```

```
str(ndcp_estimates)
```

```
## tibble [56 x 17] (S3: tbl_df/tbl/data.frame)
```

```
##  $ State                        : chr [1:56] "Alabama" "Alaska" "Arizona" "Arkansas" ...
##  $ Infant Center-Based 2018     : num [1:56] 6728 14125 10942 6344 17025 ...
##  $ Infant Home-Based 2018       : num [1:56] 5382 9150 5617 5219 11470 ...
##  $ Toddler Center-Based 2018    : num [1:56] 6728 12809 9761 6085 12085 ...
##  $ Toddler Home-Based 2018      : num [1:56] 5423 8430 5617 4913 10451 ...
##  $ Preschool Center-Based 2018 : num [1:56] 6101 10594 8428 5501 12085 ...
##  $ Preschool Home-Based 2018    : num [1:56] 5357 7800 5482 4665 10451 ...
##  $ School-Age Center-Based 2018: num [1:56] 5561 9444 7349 4852 10158 ...
##  $ School-Age Home-Based 2018   : num [1:56] 5063 6609 5217 4454 8657 ...
##  $ Infant Center-Based 2023     : num [1:56] 7919 16626 12879 7467 20039 ...
##  $ Infant Home-Based 2023       : num [1:56] 6335 10770 6611 6143 13501 ...
##  $ Toddler Center-Based 2023    : num [1:56] 7919 15077 11489 7162 14225 ...
##  $ Toddler Home-Based 2023      : num [1:56] 6384 9923 6611 5782 12301 ...
##  $ Preschool Center-Based 2023 : num [1:56] 7182 12470 9920 6475 14225 ...
##  $ Preschool Home-Based 2023    : num [1:56] 6306 9181 6453 5491 12301 ...
##  $ School-Age Center-Based 2023: num [1:56] 6546 11116 8650 5712 11957 ...
##  $ School-Age Home-Based 2023   : num [1:56] 5959 7779 6140 5243 10190 ...
```

```
# Display first few rows of the datasets
head(fl_household_demo)
```

```
## # A tibble: 6 x 10
##   `Total Population` `Total Household` `Type of Household`
##                <dbl>             <dbl> <chr>
## 1           21339762           8157420 Family
## 2                 NA                NA Non-Family
## 3                 NA                NA <NA>
## 4                 NA                NA <NA>
## 5                 NA                NA <NA>
## 6                 NA                NA <NA>
## # i 7 more variables: `Total of Type of Household` <dbl>,
## #   `Children Households` <chr>, `Total of Children Households` <dbl>,
## #   `Education Level` <chr>, `Total Education Level` <dbl>,
## #   `Marital Status` <chr>, `Marital Status Total` <dbl>
```

```
head(non_vital_rates)
```

```
## # A tibble: 6 x 11
##   County   `2022` `2021` `2020` `2019` `2018` `2017` `2016` `2015` `2014` `2013`
##   <chr>     <dbl>  <dbl>  <dbl>  <dbl>  <dbl>  <dbl>  <dbl>  <dbl>  <dbl>  <dbl>
## 1 Florida   67917  61777  57703  55660  53267  50883  48900  47507  47212  46956
## 2 Alachua   57566  53314  50089  49689  49078  45478  44702  43073  42045  42149
## 3 Baker     67872  63860  62299  63275  61769  59506  53327  47121  46865  49236
## 4 Bay       65999  60473  56483  54316  51829  50283  48577  47368  47274  47461
## 5 Bradford  54759  48803  43580  45921  46197  46106  43373  41606  40481  40259
## 6 Brevard   71308  63632  59359  56775  54359  51536  49914  48925  48483  48039
```

```
head(ndcp_estimates)
```

```
## # A tibble: 6 x 17
##   State     Infant Center-Based ~1 Infant Home-Based 20~2 Toddler Center-Based~3
##   <chr>                      <dbl>                  <dbl>                   <dbl>
## 1 Alabama                    6728.                  5382.                   6728.
## 2 Alaska                    14125.                  9150.                  12809.
## 3 Arizona                   10942.                  5617.                   9761.
## 4 Arkansas                   6344.                  5219.                   6085.
```

```
## 5 Californ~                 17025.                 11470.                 12085.
## 6 Colorado                     NA                     NA                     NA
## # i abbreviated names: 1: `Infant Center-Based 2018`,
## #   2: `Infant Home-Based 2018`, 3: `Toddler Center-Based 2018`
## # i 13 more variables: `Toddler Home-Based 2018` <dbl>,
## #   `Preschool Center-Based 2018` <dbl>, `Preschool Home-Based 2018` <dbl>,
## #   `School-Age Center-Based 2018` <dbl>, `School-Age Home-Based 2018` <dbl>,
## #   `Infant Center-Based 2023` <dbl>, `Infant Home-Based 2023` <dbl>,
## #   `Toddler Center-Based 2023` <dbl>, `Toddler Home-Based 2023` <dbl>, ...
```

```
# Summary of the datasets
summary(fl_household_demo)
```

```
##  Total Population   Total Household   Type of Household
##  Min.   :21339762   Min.   :8157420   Length:6
##  1st Qu.:21339762   1st Qu.:8157420   Class :character
##  Median :21339762   Median :8157420   Mode  :character
##  Mean   :21339762   Mean   :8157420
##  3rd Qu.:21339762   3rd Qu.:8157420
##  Max.   :21339762   Max.   :8157420
##  NA's   :5          NA's   :5
##  Total of Type of Household Children Households Total of Children Households
##  Min.   :2882929            Length:6            Min.   :2196679
##  1st Qu.:3480820            Class :character    1st Qu.:3137694
##  Median :4078710            Mode  :character    Median :4078710
##  Mean   :4078710                                Mean   :4078710
##  3rd Qu.:4676600                                3rd Qu.:5019726
##  Max.   :5274491                                Max.   :5960741
##  NA's   :4                                      NA's   :4
##  Education Level    Total Education Level Marital Status
##  Length:6           Min.   : 707827       Length:6
##  Class :character   1st Qu.:1740493       Class :character
##  Mode  :character   Median :2513302       Mode  :character
##                     Mean   :3009414
##                     3rd Qu.:4210526
##                     Max.   :6062899
##
##  Marital Status Total
##  Min.   : 358544
##  1st Qu.:1169892
##  Median :2297142
##  Mean   :3635866
##  3rd Qu.:5606403
##  Max.   :8747349
##  NA's   :1
```

```
summary(non_vital_rates)
```

```
##     County               2022             2021             2020
##  Length:68          Min.   : 37221   Min.   :38088   Min.   :35240
##  Class :character   1st Qu.: 51299   1st Qu.:46989   1st Qu.:43676
##  Mode  :character   Median : 62620   Median :57072   Median :53227
##                     Mean   : 61876   Mean   :56440   Mean   :53081
##                     3rd Qu.: 70401   3rd Qu.:63936   3rd Qu.:60114
##                     Max.   :100020   Max.   :88794   Max.   :83803
```

```
##       2019            2018            2017            2016
##  Min.   :35438   Min.   :34583   Min.   :31816   Min.   :29806
##  1st Qu.:41670   1st Qu.:40638   1st Qu.:39062   1st Qu.:37887
##  Median :51131   Median :49152   Median :46822   Median :44463
##  Mean   :51354   Mean   :49108   Mean   :47199   Mean   :45260
##  3rd Qu.:58133   3rd Qu.:54825   3rd Qu.:52391   3rd Qu.:50663
##  Max.   :82252   Max.   :77323   Max.   :73640   Max.   :69523
##       2015            2014            2013
##  Min.   :31715   Min.   :32714   Min.   :32497
##  1st Qu.:36651   1st Qu.:37588   1st Qu.:37982
##  Median :43407   Median :43081   Median :43526
##  Mean   :44097   Mean   :43957   Mean   :43764
##  3rd Qu.:48634   3rd Qu.:48052   3rd Qu.:48110
##  Max.   :66194   Max.   :65575   Max.   :64876
```

```
summary(ndcp_estimates)
```

```
##     State           Infant Center-Based 2018 Infant Home-Based 2018
##  Length:56          Min.   : 4131            Min.   : 3598
##  Class :character   1st Qu.: 8212            1st Qu.: 6401
##  Mode  :character   Median :10821            Median : 7624
##                     Mean   :10925            Mean   : 7801
##                     3rd Qu.:12685            3rd Qu.: 8981
##                     Max.   :19703            Max.   :13193
##                     NA's   :9                NA's   :9
##  Toddler Center-Based 2018 Toddler Home-Based 2018 Preschool Center-Based 2018
##  Min.   : 3911             Min.   : 3596          Min.   : 3911
##  1st Qu.: 7407             1st Qu.: 6069          1st Qu.: 7017
##  Median : 9761             Median : 6903          Median : 8878
##  Mean   : 9846             Mean   : 7371          Mean   : 8866
##  3rd Qu.:11519             3rd Qu.: 8675          3rd Qu.:10609
##  Max.   :18000             Max.   :11984          Max.   :14079
##  NA's   :9                 NA's   :9              NA's   :9
##  Preschool Home-Based 2018 School-Age Center-Based 2018
##  Min.   : 3107             Min.   : 1800
##  1st Qu.: 5997             1st Qu.: 5476
##  Median : 6820             Median : 6604
##  Mean   : 7155             Mean   : 6873
##  3rd Qu.: 8224             3rd Qu.: 8852
##  Max.   :11328             Max.   :12239
##  NA's   :9                 NA's   :9
##  School-Age Home-Based 2018 Infant Center-Based 2023 Infant Home-Based 2023
##  Min.   : 2522              Min.   : 4862            Min.   : 4234
##  1st Qu.: 5069              1st Qu.: 9666            1st Qu.: 7534
##  Median : 6200              Median :12737            Median : 8974
##  Mean   : 6248              Mean   :12859            Mean   : 9182
##  3rd Qu.: 7247              3rd Qu.:14931            3rd Qu.:10571
##  Max.   :10499              Max.   :23191            Max.   :15529
##  NA's   :9                  NA's   :9                NA's   :9
##  Toddler Center-Based 2023 Toddler Home-Based 2023 Preschool Center-Based 2023
##  Min.   : 4603             Min.   : 4233          Min.   : 4603
##  1st Qu.: 8719             1st Qu.: 7144          1st Qu.: 8259
##  Median :11489             Median : 8125          Median :10450
##  Mean   :11589             Mean   : 8676          Mean   :10436
##  3rd Qu.:13558             3rd Qu.:10211          3rd Qu.:12488
```

```
##  Max.   :21187             Max.   :14106           Max.   :16572
##  NA's   :9                 NA's   :9               NA's   :9
##  Preschool Home-Based 2023 School-Age Center-Based 2023
##  Min.   : 3657             Min.   : 2119
##  1st Qu.: 7059             1st Qu.: 6445
##  Median : 8027             Median : 7773
##  Mean   : 8422             Mean   : 8090
##  3rd Qu.: 9680             3rd Qu.:10419
##  Max.   :13333             Max.   :14406
##  NA's   :9                 NA's   :9
##  School-Age Home-Based 2023
##  Min.   : 2969
##  1st Qu.: 5966
##  Median : 7298
##  Mean   : 7355
##  3rd Qu.: 8530
##  Max.   :12358
##  NA's   :9
```

```
# Data Preparation and Cleaning Steps
# Florida Household Demographic
fl_household_demo <- fl_household_demo %>%
  filter(!is.na(`Total Population`)) %>%
  mutate(
    `Total Population` = as.numeric(`Total Population`),
    `Total Household` = as.numeric(`Total Household`)
  )

# NonVitalRateOnlyInd_TenYrsReport
# Convert specific columns to numeric and handle non-numeric values
numeric_cols_non_vital <- colnames(non_vital_rates)[2:11]  # Assuming columns 2 to 11 are the year colu
non_vital_rates <- non_vital_rates %>%
  mutate(across(all_of(numeric_cols_non_vital), ~ as.numeric(as.character(.))))

# NDCP-State-Level-Estimates-2018-2023
# Convert specific columns to numeric and handle non-numeric values
numeric_cols_ndcp <- colnames(ndcp_estimates)[2:17]  # Assuming columns 2 to 17 are the numeric cost co
ndcp_estimates <- ndcp_estimates %>%
  mutate(across(all_of(numeric_cols_ndcp), ~ as.numeric(as.character(.))))

# Display the cleaned datasets
head(fl_household_demo)
```

```
## # A tibble: 1 x 10
##   `Total Population` `Total Household` `Type of Household`
##                <dbl>             <dbl> <chr>
## 1           21339762           8157420 Family
## # i 7 more variables: `Total of Type of Household` <dbl>,
## #   `Children Households` <chr>, `Total of Children Households` <dbl>,
## #   `Education Level` <chr>, `Total Education Level` <dbl>,
## #   `Marital Status` <chr>, `Marital Status Total` <dbl>
```

```
head(non_vital_rates)
```

```
## # A tibble: 6 x 11
##   County   `2022` `2021` `2020` `2019` `2018` `2017` `2016` `2015` `2014` `2013`
```

```
##   <chr>     <dbl>  <dbl>  <dbl>  <dbl>  <dbl>  <dbl>  <dbl>  <dbl>  <dbl>  <dbl>
## 1 Florida   67917  61777  57703  55660  53267  50883  48900  47507  47212  46956
## 2 Alachua   57566  53314  50089  49689  49078  45478  44702  43073  42045  42149
## 3 Baker     67872  63860  62299  63275  61769  59506  53327  47121  46865  49236
## 4 Bay       65999  60473  56483  54316  51829  50283  48577  47368  47274  47461
## 5 Bradford  54759  48803  43580  45921  46197  46106  43373  41606  40481  40259
## 6 Brevard   71308  63632  59359  56775  54359  51536  49914  48925  48483  48039
```

```
head(ndcp_estimates)
```

```
## # A tibble: 6 x 17
##   State     Infant Center-Based ~1 Infant Home-Based 20~2 Toddler Center-Based~3
##   <chr>                      <dbl>                  <dbl>                  <dbl>
## 1 Alabama                    6728.                  5382.                  6728.
## 2 Alaska                    14125.                  9150.                 12809.
## 3 Arizona                   10942.                  5617.                  9761.
## 4 Arkansas                   6344.                  5219.                  6085.
## 5 Californ~                 17025.                 11470.                 12085.
## 6 Colorado                     NA                     NA                     NA
## # i abbreviated names: 1: `Infant Center-Based 2018`,
## #   2: `Infant Home-Based 2018`, 3: `Toddler Center-Based 2018`
## # i 13 more variables: `Toddler Home-Based 2018` <dbl>,
## #   `Preschool Center-Based 2018` <dbl>, `Preschool Home-Based 2018` <dbl>,
## #   `School-Age Center-Based 2018` <dbl>, `School-Age Home-Based 2018` <dbl>,
## #   `Infant Center-Based 2023` <dbl>, `Infant Home-Based 2023` <dbl>,
## #   `Toddler Center-Based 2023` <dbl>, `Toddler Home-Based 2023` <dbl>, ...
```

```
# Example Analysis
# Calculate summary statistics
household_summary <- fl_household_demo %>%
  summarise(
    total_population = sum(`Total Population`, na.rm = TRUE),
    total_household = sum(`Total Household`, na.rm = TRUE)
  )

income_summary <- non_vital_rates %>%
  summarise(across(all_of(numeric_cols_non_vital), mean, na.rm = TRUE))
```

```
## Warning: There was 1 warning in `summarise()`.
## i In argument: `across(all_of(numeric_cols_non_vital), mean, na.rm = TRUE)`.
## Caused by warning:
## ! The `...` argument of `across()` is deprecated as of dplyr 1.1.0.
## Supply arguments directly to `.fns` through an anonymous function instead.
##
##   # Previously
##   across(a:b, mean, na.rm = TRUE)
##
##   # Now
##   across(a:b, \(x) mean(x, na.rm = TRUE))
```

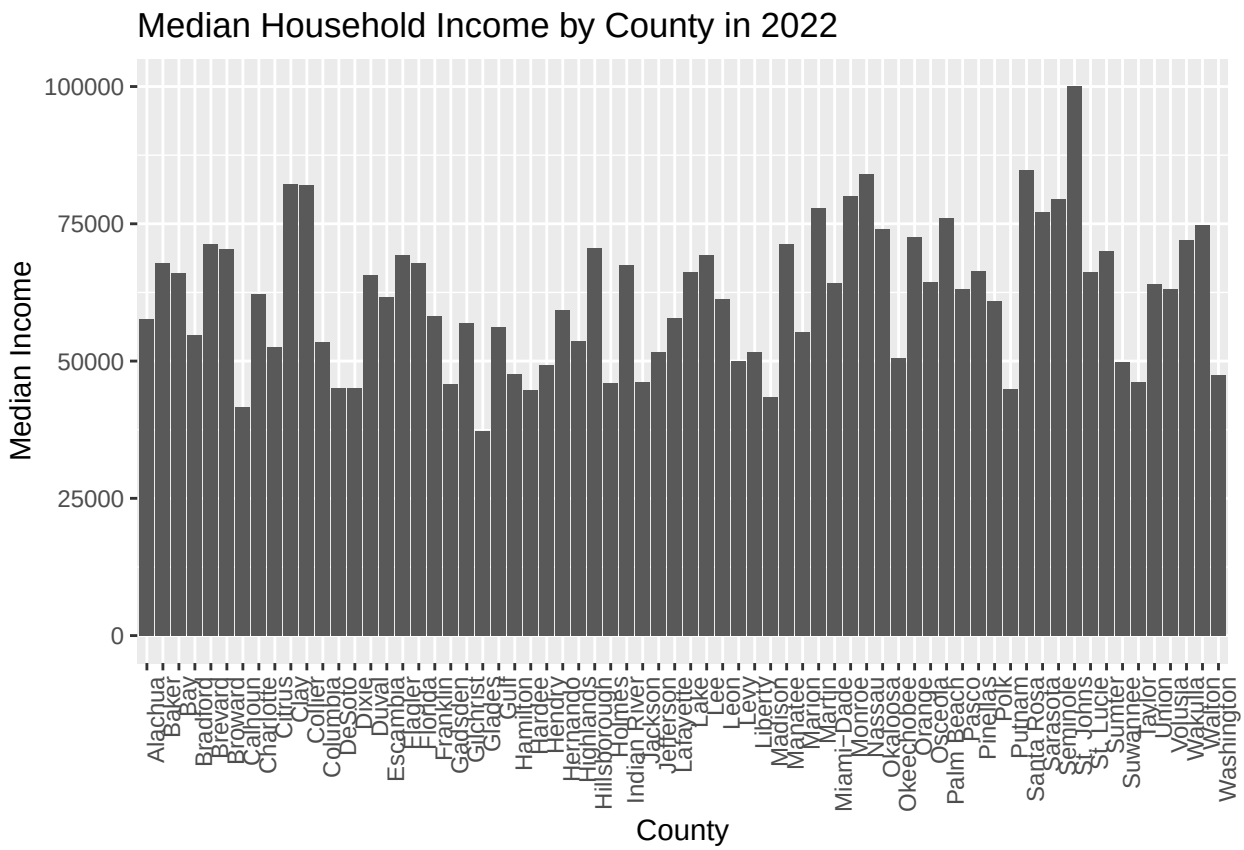
```
childcare_cost_summary <- ndcp_estimates %>%
  summarise(across(all_of(numeric_cols_ndcp), mean, na.rm = TRUE))

# Create example plots
# Filter out rows with missing values before plotting
non_vital_rates_plot <- non_vital_rates %>%
```

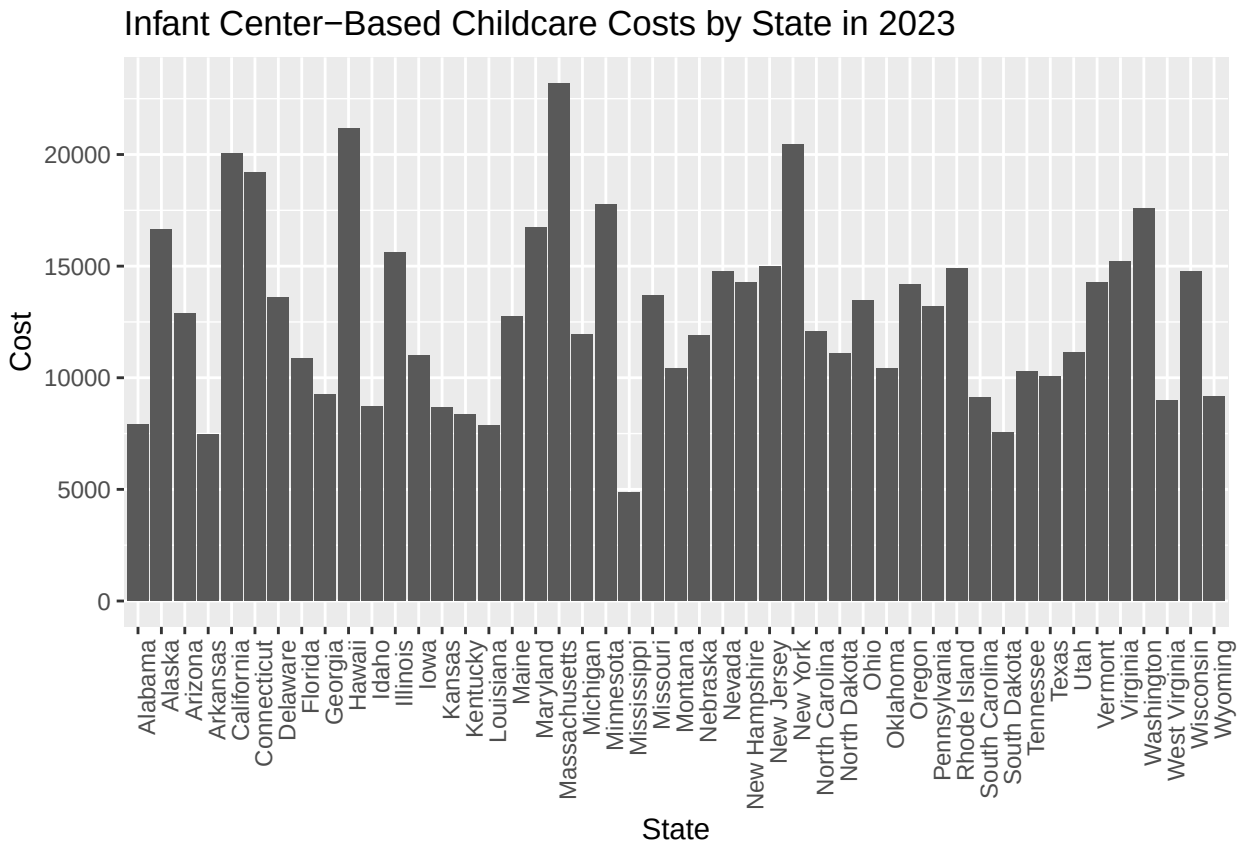
```
  filter(!is.na(`2022`))

ndcp_estimates_plot <- ndcp_estimates %>%
  filter(!is.na(`Infant Center-Based 2023`))

ggplot(non_vital_rates_plot, aes(x = County, y = `2022`)) +
  geom_bar(stat = "identity") +
  labs(title = "Median Household Income by County in 2022", x = "County", y = "Median Income") +
  theme(axis.text.x = element_text(angle = 90, hjust = 1))
```




```
ggplot(ndcp_estimates_plot, aes(x = State, y = `Infant Center-Based 2023`)) +
  geom_bar(stat = "identity") +
  labs(title = "Infant Center-Based Childcare Costs by State in 2023", x = "State", y = "Cost") +
  theme(axis.text.x = element_text(angle = 90, hjust = 1))
```




```
# Additional Steps
# Identify and fill missing values
# Explore relationships between household income, demographics, and childcare costs
# Potentially incorporate machine learning techniques for predictive analysis
```