Project 3

Name: Justin Campbell, UT eID: jsc4348

This is the dataset used in this project:

```
# Load in data from repository

income_dist <- readr::read_csv('https://raw.githubusercontent.com/rfordatascience/tidytuesday/master/da
income_dist
```

```
## # A tibble: 2,916 x 9
##     year race  number income_median income_med_moe income_mean income_mean_moe
##    <dbl> <chr>  <dbl>         <dbl>          <dbl>       <dbl>           <dbl>
##  1  2019 All ~ 1.28e8         68703            904       98088            1042
##  2  2019 All ~ 1.28e8         68703            904       98088            1042
##  3  2019 All ~ 1.28e8         68703            904       98088            1042
##  4  2019 All ~ 1.28e8         68703            904       98088            1042
##  5  2019 All ~ 1.28e8         68703            904       98088            1042
##  6  2019 All ~ 1.28e8         68703            904       98088            1042
##  7  2019 All ~ 1.28e8         68703            904       98088            1042
##  8  2019 All ~ 1.28e8         68703            904       98088            1042
##  9  2019 All ~ 1.28e8         68703            904       98088            1042
## 10  2018 All ~ 1.29e8         64324            704       91652             914
## # ... with 2,906 more rows, and 2 more variables: income_bracket <chr>,
## #   income_distribution <dbl>
```

Link to the dataset: *https://github.com/rfordatascience/tidytuesday/blob/master/data/2021/2021-02-09/income_distribution.csv*

Part 1

Question: *How has the median household income changed over the time period of 1987 - 2019 across each of the four individual racial groups (Asian, White, Black, and Hispanic)? How is the income distribution How is the income distributed across each of the income brackets for each of these racial groups for the year 2019? Lastly, visualize the variation in the income distribution across each of the four individual racial groups over the years 1967-2019.*

Introduction: *In this project, the author elected to operate on a dataset containing quantitative data regarding yearly household income metrics in the United States. This dataset, entitled, "income_distribution.csv", is accessible from the following webpage: "https://github.com/rfordatascience/tidytuesday/blob/master/data/2021/2021-02-09/income_distribution.csvTo expand, the elements of the dataset are broken down into median household income, median household income margin of error, mean household income, mean household income margin of error, income bracket, and income distribution according to different racial groups (Asian, White, Black, Hispanic, and a combination thereof), and over the time period of 1967 - 2019.*

In Part 1, the author first aimed to study how the median household income has changed with time across the four major racial groups contained in the dataset, namely, Asian, Black, White, and Hispanic. The author was not only interested in relationships within the individual groups, but also, across the four groups as a whole, with an intention of connecting trends in the dataset to societal and economical observations. To accomplish this, the quantitative data from the "year", and "median_income" columns, and the qualitative *data from the "race" column will be needed. The author will then explore the second question in the problem statement which involves studying how household income was distributed across the income bracket groups for each racial category in the year 2019. To answer this question, the data from the columns "race", "income_bracket", "year", and "income_distribution" will be needed. Lastly, the author will investigate how the income distribution values vary in total frequency across each of the four primary racial groups over the time period of 1967-2019. This will be accomplished using the data from columns, "year", "race", "income_distribution".*

Approach: *For the first question in part 1, the author seeks to use time series scatter plots with linear regression to visualize how the median household income has changed over time across the different racial groups. The reasoning for using time series scatter plots is that they are ideal for visualizing how quantitative data changes over time. By fitting a linear best fit line to each cluster of data (pertaining to each of the four racial groups), the author can best visualize and compare the median household income trends over time. The author will start by filtering the dataframe to include median household income values across the four primary racial groups over the period of 1987 - 2019 using "filter(year >= 1987, race =c())" framework. Then, the following line of code will be used to remove all columns with exception to the "year", "race", and "median_income", "select(year, race, median_income)". The last step in pre-processing the data before the analysis will involve removing duplicate entries in the dataset using the "unique()" function. For the analysis step, the following code framework will be used to generate a scatter plot with four clusters, and associated linear best fit lines "ggplot(dataframe, aes(year, income_median, color = race)) + geom_point() + geom_smooth()"*

For the second question in part 1, the author plans to use pie chart subplots to depict the distribution of household income across each of the income brackets for each racial group. This was determined to be a reasonable approach to visualizing the distributions given that pie charts allow for users to display qualitative data (such as that of percentages assigned to different income bracket categories) as proportions of a whole. It should be noted that the income_distribution values sum to 100% for each resultant data frame. Additionally, pie charts are visually appealing for small datasets such as that of the reduced data frames containing the income distribution and income bracket values.To pre-process the data, for each of the four racial groups, the author will first filter the data to only include values from the year 2019, and the associated race. Then, the income_bracket and income_distribution values will be values will be preserved using "select()". From here, a pie_chart data frame will be generated for each of the racial groups using the framework on slide 36 of "Visualizing proportions.pdf". Then, the pie charts will be generated using these data frames according to the template shown on slide 38 of the same document. For the third question in part 1, the author will use a histogram to visualize the frequency distribution of income for each racial group. In particular, for each racial group, the "filter()" method will be used to retain only data corresponding to that particular race from the original data set, and over the years 1967-2019. Then, the "group_by()" method will be used to group the data according to the income distribution, and the "n()" method in the summarize() function will be used to generate a new column with the frequency of each of the income distribution values. These resultant summary tables will then be returned to the console. "ggplot()" and "geom_histogram()" will then be used to generate the histograms with a readable bin spacing. The utility of using a histogram to visualize the data in this question is such that it enables the author to depict the distribution of quantitative data (such as income distribution values in the dataset). Thus, a histogram is a good choice for this application.

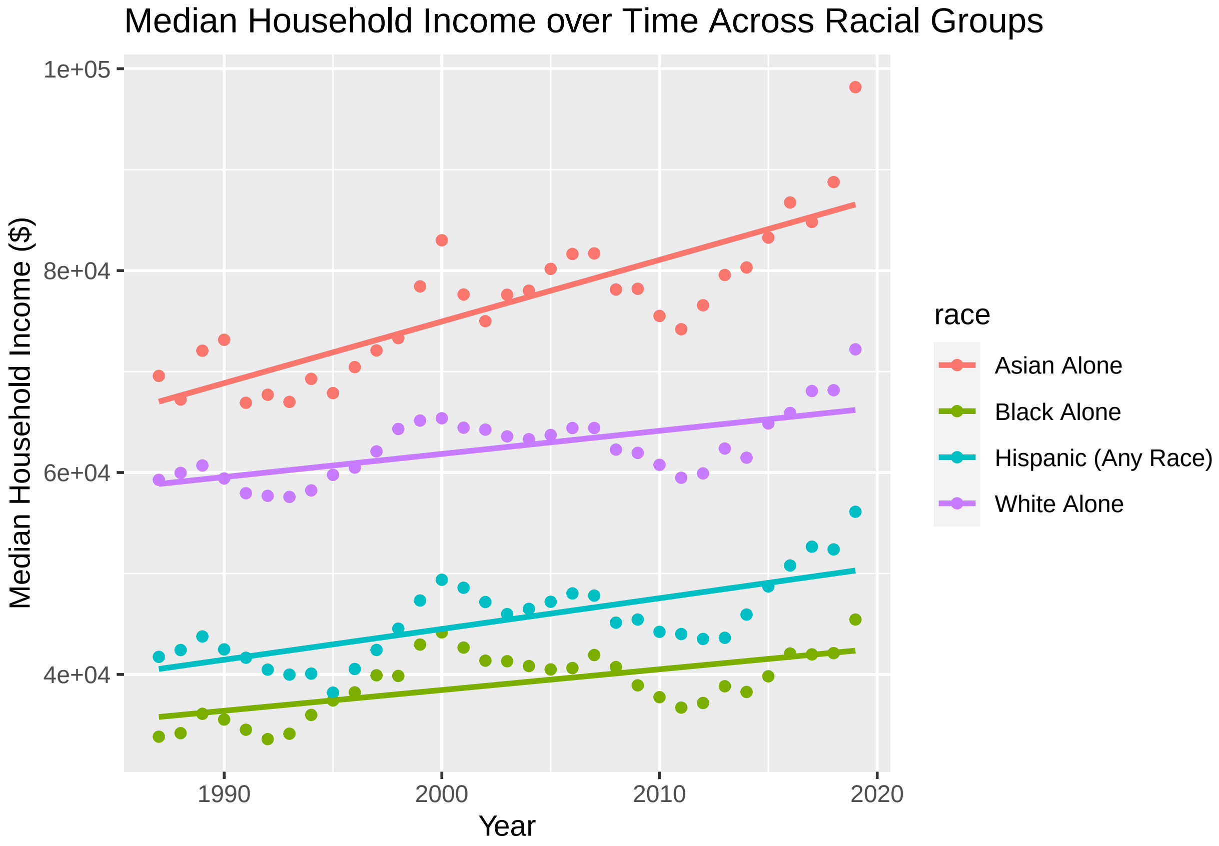
Analysis:

```r
# Filter data frame to include median household income values across four races (Asian,
#Black, Hispanic, and White) over the time period of 1987 - 2019

income_dist_1 <- income_dist %>%
  filter(race == c("Asian Alone", "Black Alone", "Hispanic (Any Race)", "White Alone"),
         year >= 1987) %>%
  select(year, race,income_median) %>%
  unique() # Remove duplicate income_median values for each year and race

income_dist_1 # Output resultant data frame to console
```

```
## # A tibble: 132 x 3
##     year race         income_median
##    <dbl> <chr>                <dbl>
##  1  2019 White Alone          72204
##  2  2018 White Alone          68156
##  3  2017 White Alone          68076
##  4  2016 White Alone          65901
##  5  2015 White Alone          64864
##  6  2014 White Alone          61470
##  7  2013 White Alone          62378
##  8  2012 White Alone          59912
##  9  2011 White Alone          59481
## 10  2010 White Alone          60763
## # ... with 122 more rows
```

```
# Generate scatter plot depicting median income across four races from 1987-2019
# with a linear best fit line

ggplot(income_dist_1, aes(year, income_median, color = race)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) + xlab("Year") + ylab("Median Household Income ($)") +
  ggtitle("Median Household Income over Time Across Racial Groups")
```

```
## `geom_smooth()` using formula 'y ~ x'
```



```
# Filter data frame to include income distribution percentages for the "Asian Alone" racial group for t
```

```r
income_dist_2_asian <- income_dist %>%
  filter(race == c("Asian Alone"),
         year == 2019) %>%
  select(income_bracket, income_distribution) #%>%

income_dist_2_asian # Output resultant data frame to console
```

```
## # A tibble: 9 x 2
##   income_bracket       income_distribution
##   <chr>                              <dbl>
## 1 Under $15,000                        6.5
## 2 $15,000 to $24,999                   5  
## 3 $25,000 to $34,999                   5.2
## 4 $35,000 to $49,999                   8.7
## 5 $50,000 to $74,999                  12.9
## 6 $75,000 to $99,999                  12.5
## 7 $100,000 to $149,999                17.9
## 8 $150,000 to $199,999                12.5
## 9 $200,000 and over                   18.9
```

```r
# Filter data frame to include income distribution percentages for the "Black Alone" racial group for t

income_dist_2_black <- income_dist %>%
  filter(race == c("Black Alone"),
         year == 2019) %>%
  select(income_bracket, income_distribution) #%>%

income_dist_2_black # Output resultant data frame to console
```

```
## # A tibble: 9 x 2
##   income_bracket       income_distribution
##   <chr>                              <dbl>
## 1 Under $15,000                       17.2
## 2 $15,000 to $24,999                  11.5
## 3 $25,000 to $34,999                  11.4
## 4 $35,000 to $49,999                  13.7
## 5 $50,000 to $74,999                  16.8
## 6 $75,000 to $99,999                   9.8
## 7 $100,000 to $149,999                10.8
## 8 $150,000 to $199,999                 4.2
## 9 $200,000 and over                    4.6
```

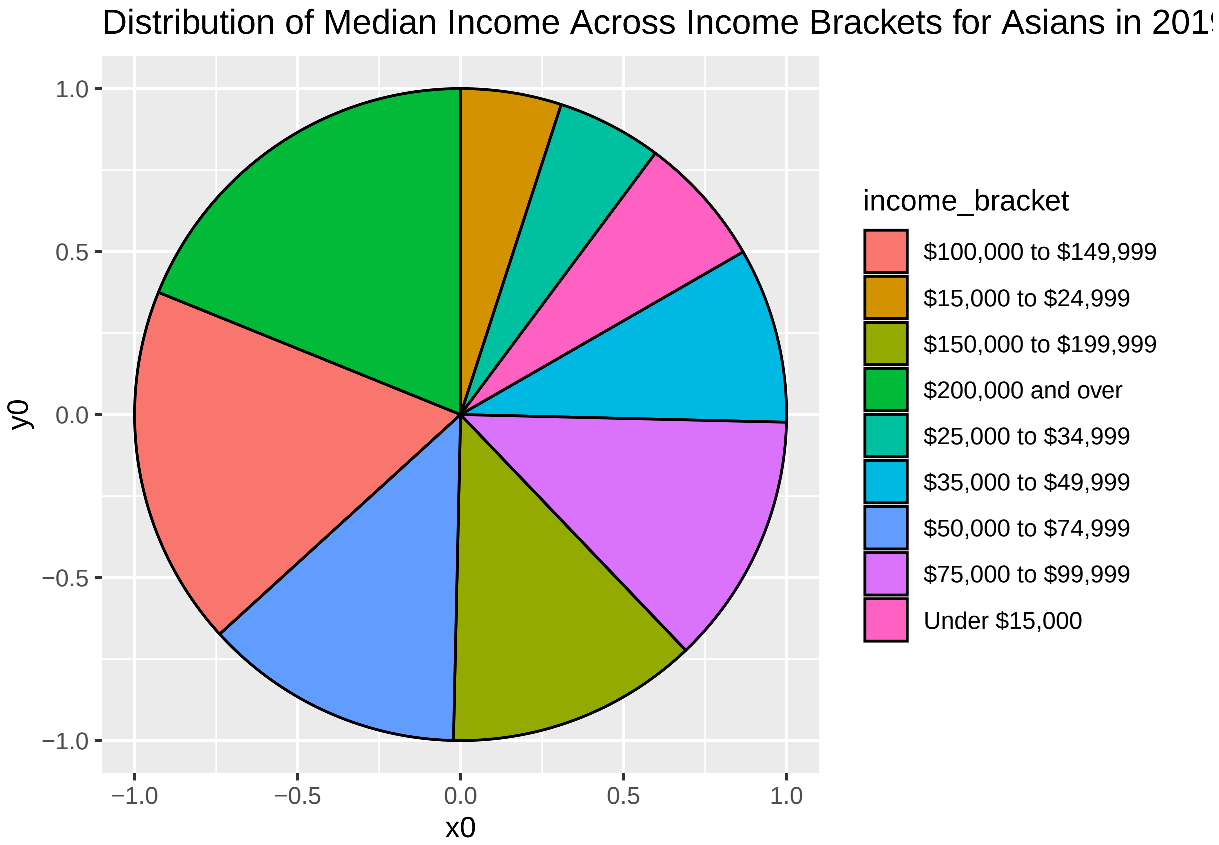
```r
# Filter data frame to include income distribution percentages for the "White Alone" racial group for t

income_dist_2_white <- income_dist %>%
  filter(race == c("White Alone"),
         year == 2019) %>%
  select(income_bracket, income_distribution) #%>%

income_dist_2_white # Output resultant data frame to console
```

```
## # A tibble: 9 x 2
##   income_bracket       income_distribution
##   <chr>                              <dbl>
## 1 Under $15,000                        7.8
## 2 $15,000 to $24,999                   7.5
## 3 $25,000 to $34,999                   8
## 4 $35,000 to $49,999                  11.5
## 5 $50,000 to $74,999                  16.7
## 6 $75,000 to $99,999                  12.7
## 7 $100,000 to $149,999                16.3
## 8 $150,000 to $199,999                 8.7
## 9 $200,000 and over                   10.8
```

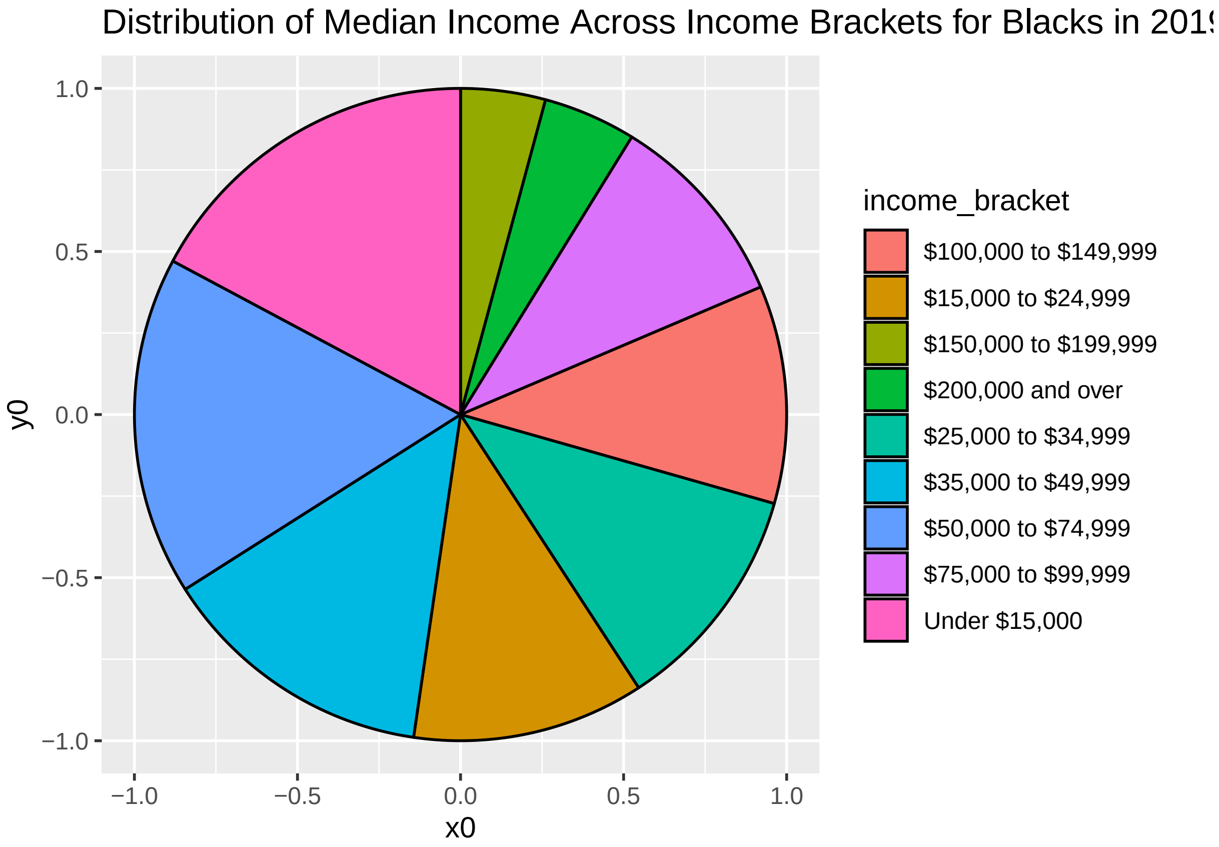
```
# Filter data frame to include income distribution percentages for the "Hispanic (Any Race)" racial gro

income_dist_2_hispanic <- income_dist %>%
  filter(race == c("Hispanic (Any Race)"),
         year == 2019) %>%
  select(income_bracket, income_distribution) #%>%

income_dist_2_hispanic # Output resultant data frame to console
```

```
## # A tibble: 9 x 2
##   income_bracket       income_distribution
##   <chr>                              <dbl>
## 1 Under $15,000                       10.7
## 2 $15,000 to $24,999                   8.8
## 3 $25,000 to $34,999                  10.5
## 4 $35,000 to $49,999                  14.1
## 5 $50,000 to $74,999                  19.5
## 6 $75,000 to $99,999                  12.2
## 7 $100,000 to $149,999                13
## 8 $150,000 to $199,999                 5.9
## 9 $200,000 and over                    5.3
```

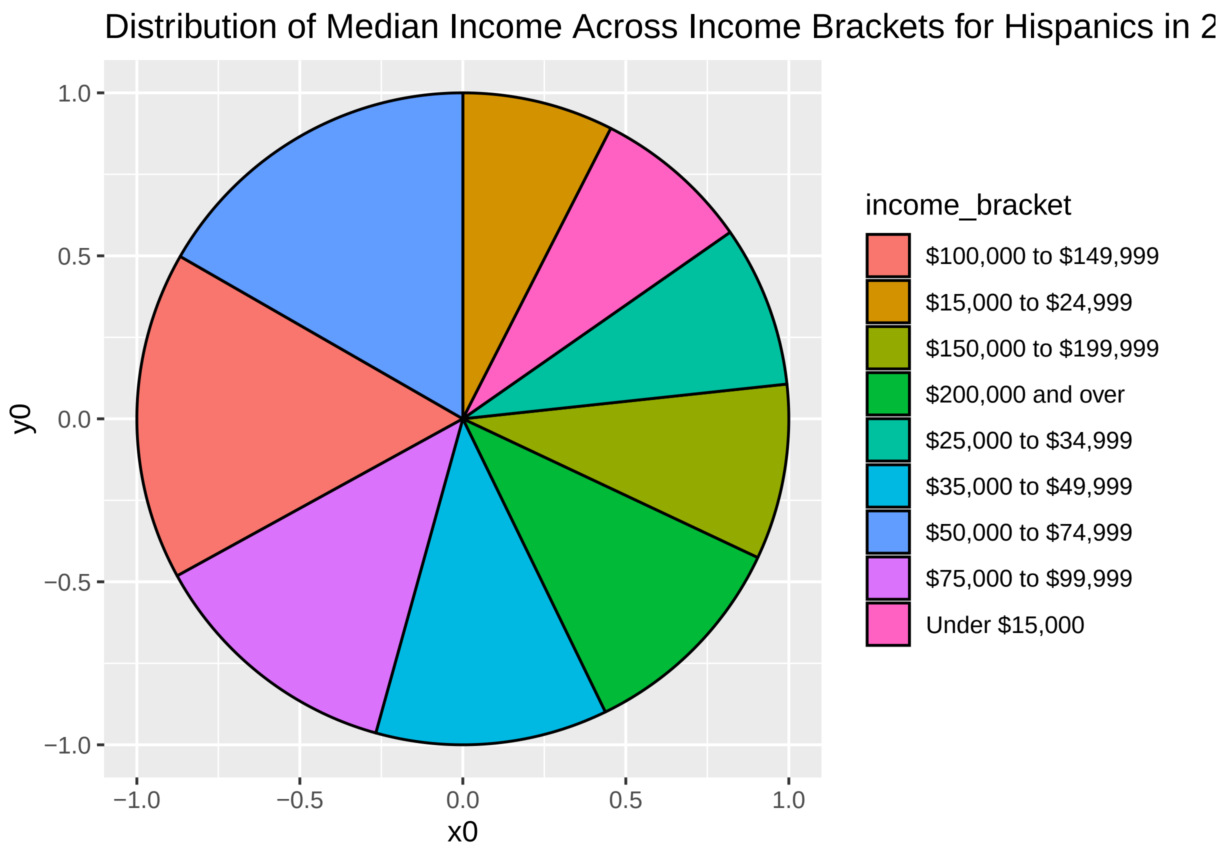
```
# Generation of Pie Chart for Asian Group

pie_data_asian <- income_dist_2_asian %>%
  arrange(income_distribution) %>% # Sort pie slices
  mutate(
    end_angle = 2*pi*cumsum(income_distribution) / sum(income_distribution), # Ending angle for pie sli
    start_angle = lag(end_angle, default = 0), # starting angle for pie slice
    mid_angle = 0.5 * (start_angle + end_angle), # middle of pie slice for text labels
    # Horizontal and Vertical Outer Label Justifications
    hjust = ifelse(mid_angle > pi, 1, 0),
    vjust = ifelse(mid_angle < pi/2 | mid_angle > 3*pi/2, 0, 1)
  )

ggplot(pie_data_asian) +
  aes(
    x0 = 0, y0 = 0, r0 = 0, r = 1,
    start = start_angle, end = end_angle,
    fill = income_bracket
  ) +
  geom_arc_bar() +
  coord_fixed() + ggtitle("Distribution of Median Income Across Income Brackets for Asians in 2019")
```



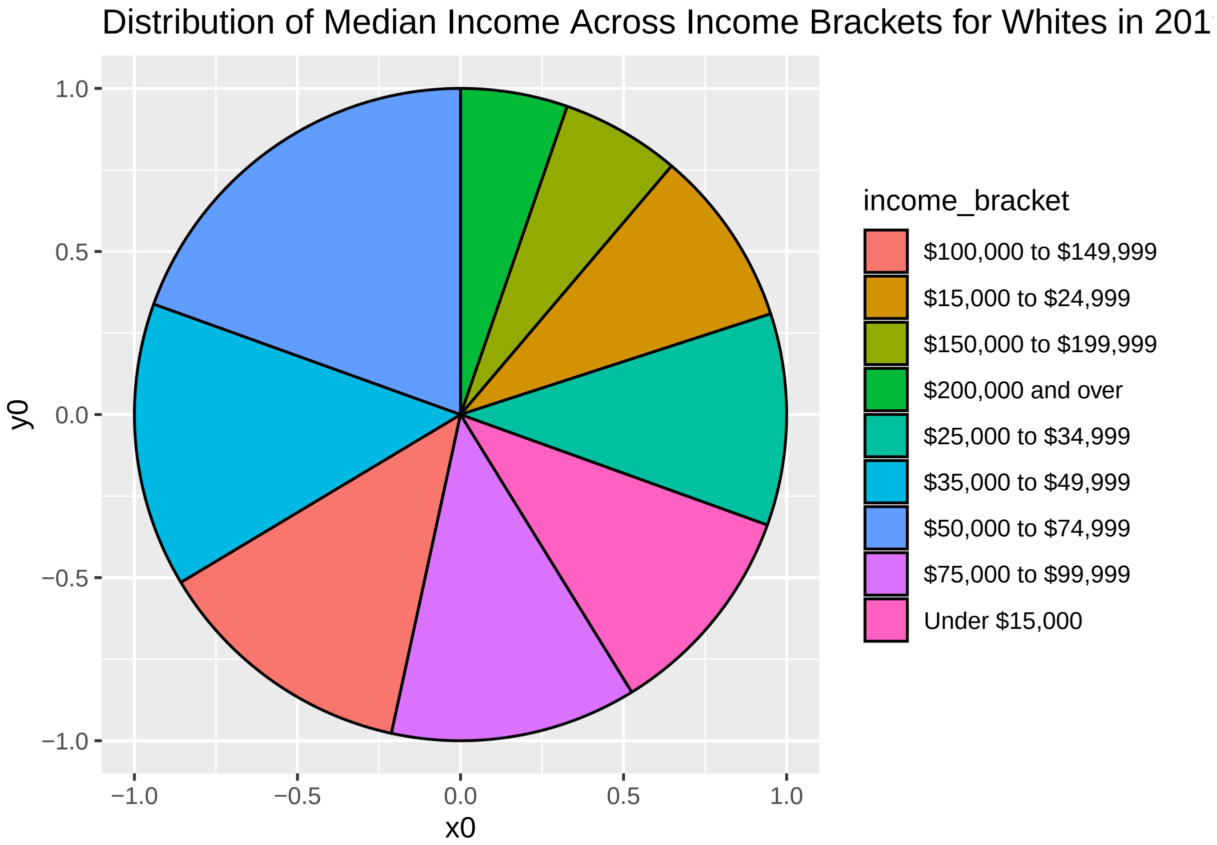
```r
# Generation of Pie Chart for Black Group

pie_data_black <- income_dist_2_black %>%
  arrange(income_distribution) %>% # Sort pie slices
  mutate(
    end_angle = 2*pi*cumsum(income_distribution) / sum(income_distribution), # Ending angle for pie sli
    start_angle = lag(end_angle, default = 0), # starting angle for pie slice
    mid_angle = 0.5 * (start_angle + end_angle), # middle of pie slice for text labels
    # Horizontal and Vertical Outer Label Justifications
    hjust = ifelse(mid_angle > pi, 1, 0),
    vjust = ifelse(mid_angle < pi/2 | mid_angle > 3*pi/2, 0, 1)
  )

ggplot(pie_data_black) +
  aes(
    x0 = 0, y0 = 0, r0 = 0, r = 1,
    start = start_angle, end = end_angle,
    fill = income_bracket
  ) +
  geom_arc_bar() +
  coord_fixed() + ggtitle("Distribution of Median Income Across Income Brackets for Blacks in 2019")
```



```r
# Generation of Pie Chart for White Group

pie_data_white <- income_dist_2_white %>%
  arrange(income_distribution) %>% # Sort pie slices
  mutate(
    end_angle = 2*pi*cumsum(income_distribution) / sum(income_distribution), # Ending angle for pie sli
    start_angle = lag(end_angle, default = 0), # starting angle for pie slice
    mid_angle = 0.5 * (start_angle + end_angle), # middle of pie slice for text labels
    # Horizontal and Vertical Outer Label Justifications
    hjust = ifelse(mid_angle > pi, 1, 0),
    vjust = ifelse(mid_angle < pi/2 | mid_angle > 3*pi/2, 0, 1)
  )

ggplot(pie_data_white) +
  aes(
    x0 = 0, y0 = 0, r0 = 0, r = 1,
    start = start_angle, end = end_angle,
    fill = income_bracket
  ) +
  geom_arc_bar() +
  coord_fixed() + ggtitle("Distribution of Median Income Across Income Brackets for Hispanics in 2019")
```



```r
# Generation of Pie Chart for Hispanic Group

pie_data_hispanic <- income_dist_2_hispanic %>%
  arrange(income_distribution) %>% # Sort pie slices
  mutate(
    end_angle = 2*pi*cumsum(income_distribution) / sum(income_distribution), # Ending angle for pie sli
    start_angle = lag(end_angle, default = 0), # starting angle for pie slice
    mid_angle = 0.5 * (start_angle + end_angle), # middle of pie slice for text labels
    # Horizontal and Vertical Outer Label Justifications
    hjust = ifelse(mid_angle > pi, 1, 0),
    vjust = ifelse(mid_angle < pi/2 | mid_angle > 3*pi/2, 0, 1)
  )

ggplot(pie_data_hispanic) +
  aes(
    x0 = 0, y0 = 0, r0 = 0, r = 1,
    start = start_angle, end = end_angle,
    fill = income_bracket
  ) +
  geom_arc_bar() +
  coord_fixed() + ggtitle("Distribution of Median Income Across Income Brackets for Whites in 2019")
```



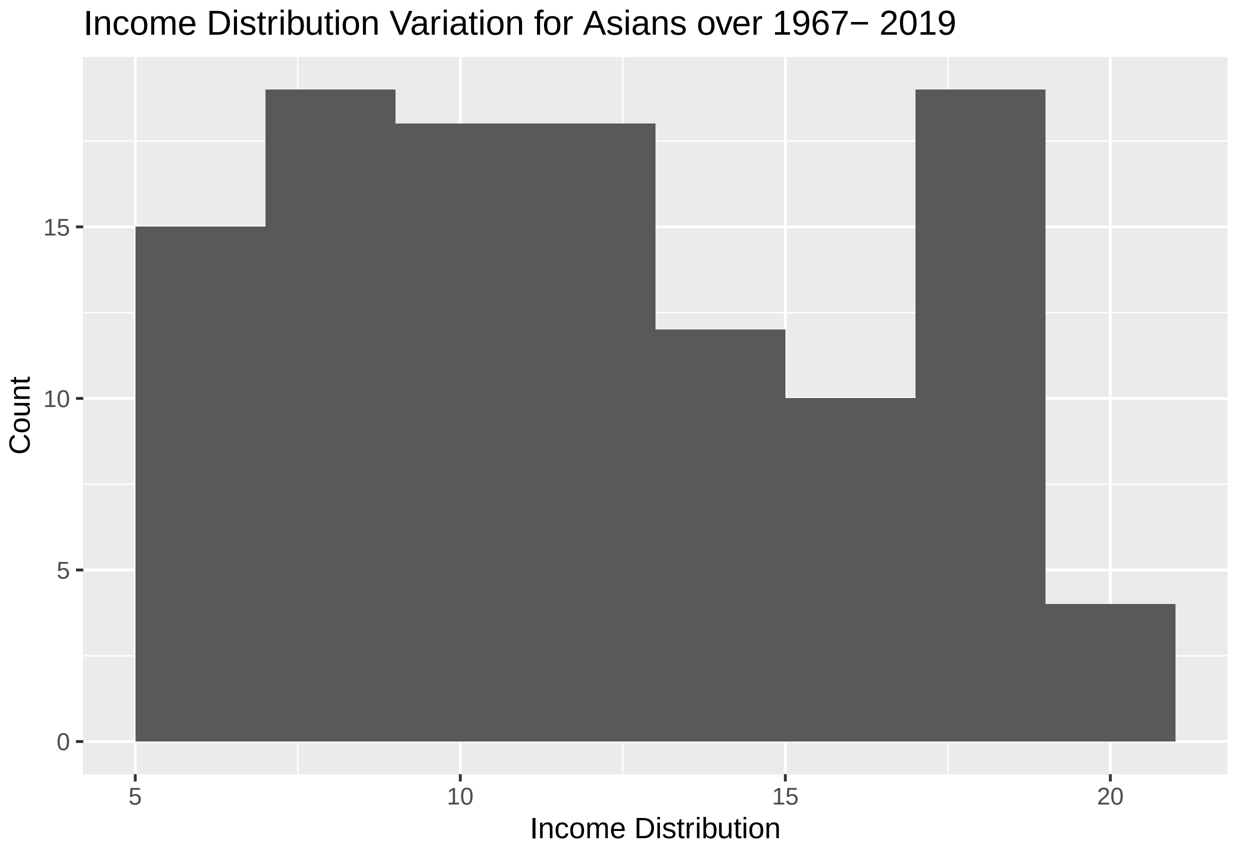
```
income_dist_3_asian <- income_dist %>%
  filter(race == "Asian Alone",
         year >= 1967) %>%
  group_by(income_distribution) %>%
  summarize(
    n = n()
  )
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
income_dist_3_asian
```

```
## # A tibble: 115 x 2
##    income_distribution     n
##                  <dbl> <int>
##  1                 5       2
##  2                 5.2     2
##  3                 5.7     1
##  4                 5.8     1
##  5                 5.9     2
##  6                 6       4
##  7                 6.1     4
##  8                 6.2     3
##  9                 6.4     5
## 10                 6.5     5
## # ... with 105 more rows
```

```
ggplot(income_dist_3_asian, aes(income_distribution)) +
  geom_histogram(binwidth = 2) + xlab("Income Distribution") + ylab("Count") + ggtitle("Income Distribu
```



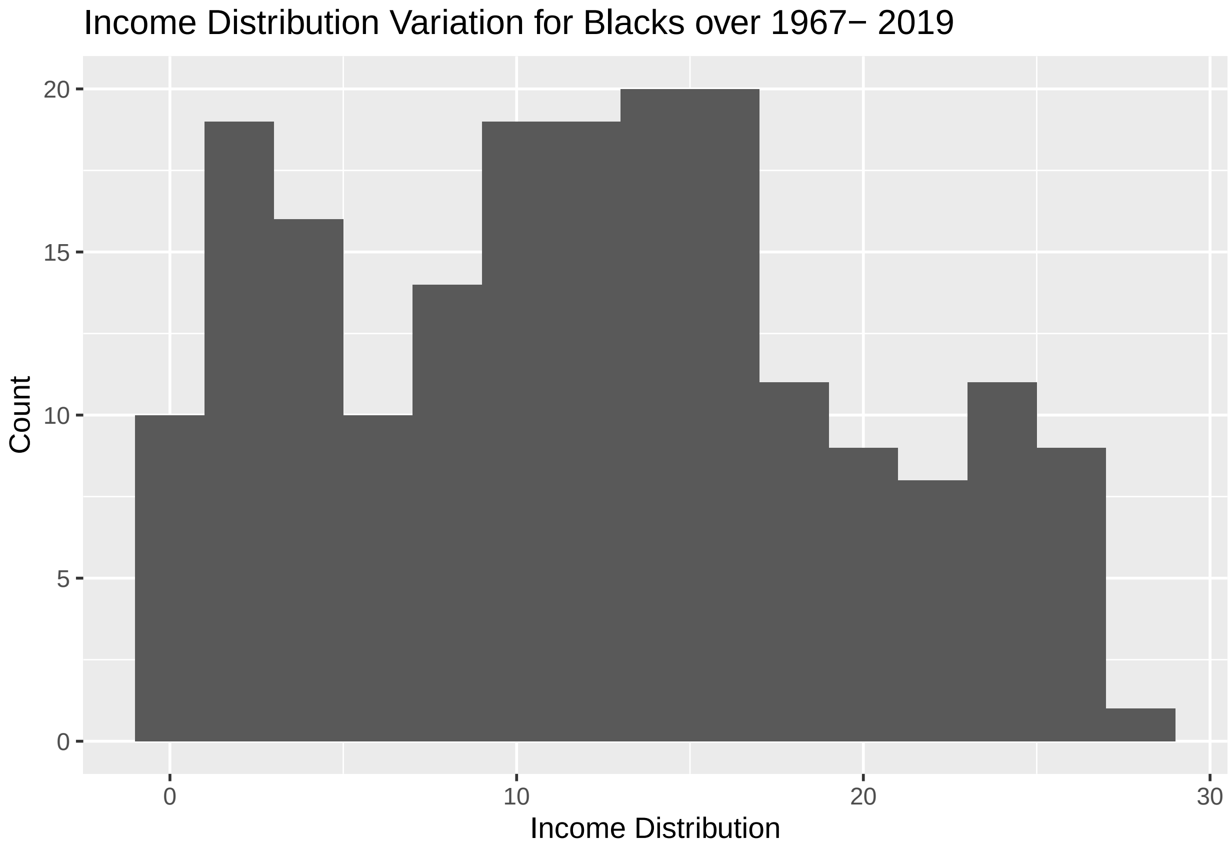
```
income_dist_3_black <- income_dist %>%
  filter(race == "Black Alone",
         year >= 1967) %>%
  group_by(income_distribution) %>%
  summarize(
    n = n()
  )
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
income_dist_3_black
```

```
## # A tibble: 196 x 2
##    income_distribution     n
##                  <dbl> <int>
##  1                 0.1     4
##  2                 0.2     5
##  3                 0.3     5
##  4                 0.4     5
##  5                 0.5     1
##  6                 0.6     6
##  7                 0.7     3
##  8                 0.8     3
##  9                 0.9     5
## 10                    1       3
## # ... with 186 more rows
```

```
ggplot(income_dist_3_black, aes(income_distribution)) +
  geom_histogram(binwidth = 2) + xlab("Income Distribution") + ylab("Count") + ggtitle("Income Distribu
```



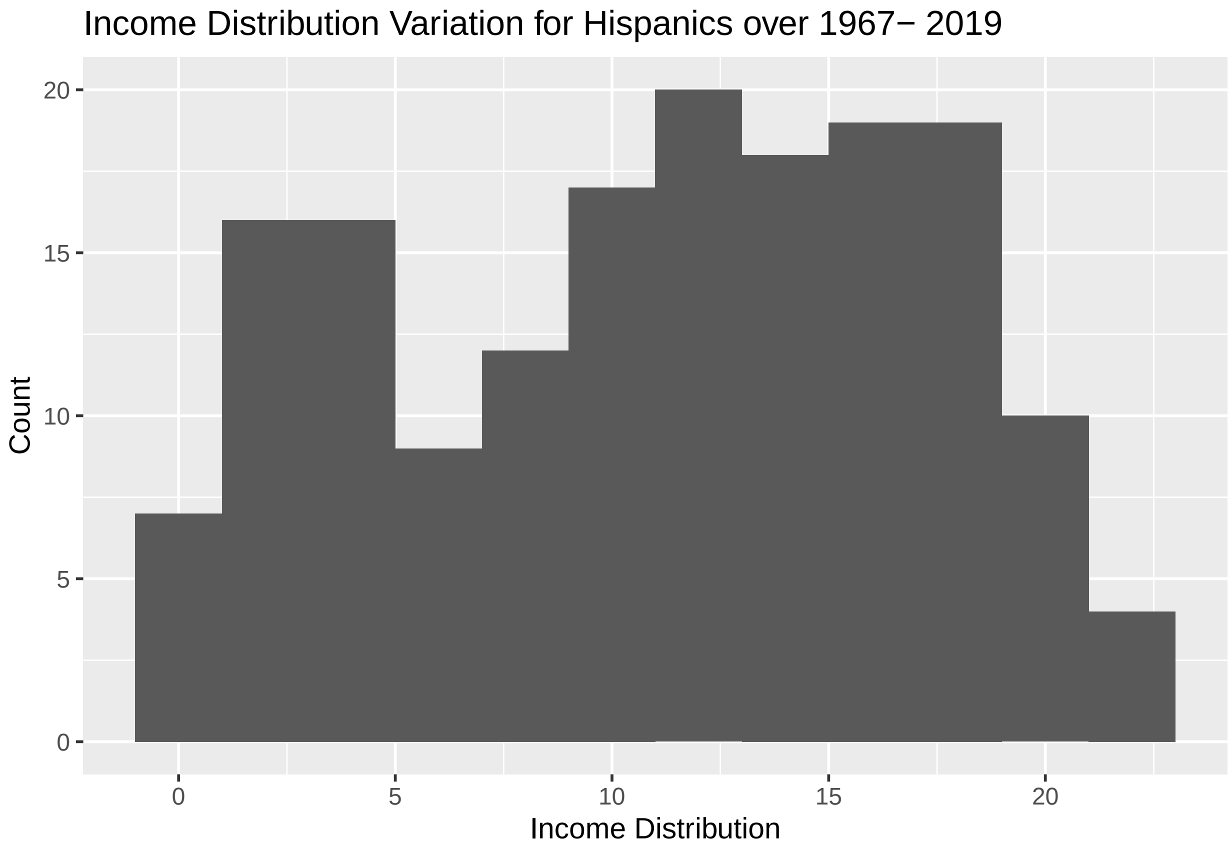
```
income_dist_3_hisp <- income_dist %>%
  filter(race == "Hispanic (Any Race)",
         year >= 1967) %>%
  group_by(income_distribution) %>%
  summarize(
    n = n()
  )
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
income_dist_3_hisp
```

```
## # A tibble: 167 x 2
##    income_distribution     n
##                  <dbl> <int>
##  1                 0.3     1
##  2                 0.4     2
##  3                 0.5     3
##  4                 0.6     2
##  5                 0.7     2
##  6                 0.8     3
##  7                 0.9     5
## 8                  1.1     4
## 9                  1.3     1
## 10                 1.4     6
## # ... with 157 more rows
```

```
ggplot(income_dist_3_hisp, aes(income_distribution)) +
  geom_histogram(binwidth = 2) + xlab("Income Distribution") + ylab("Count") + ggtitle("Income Distribu
```



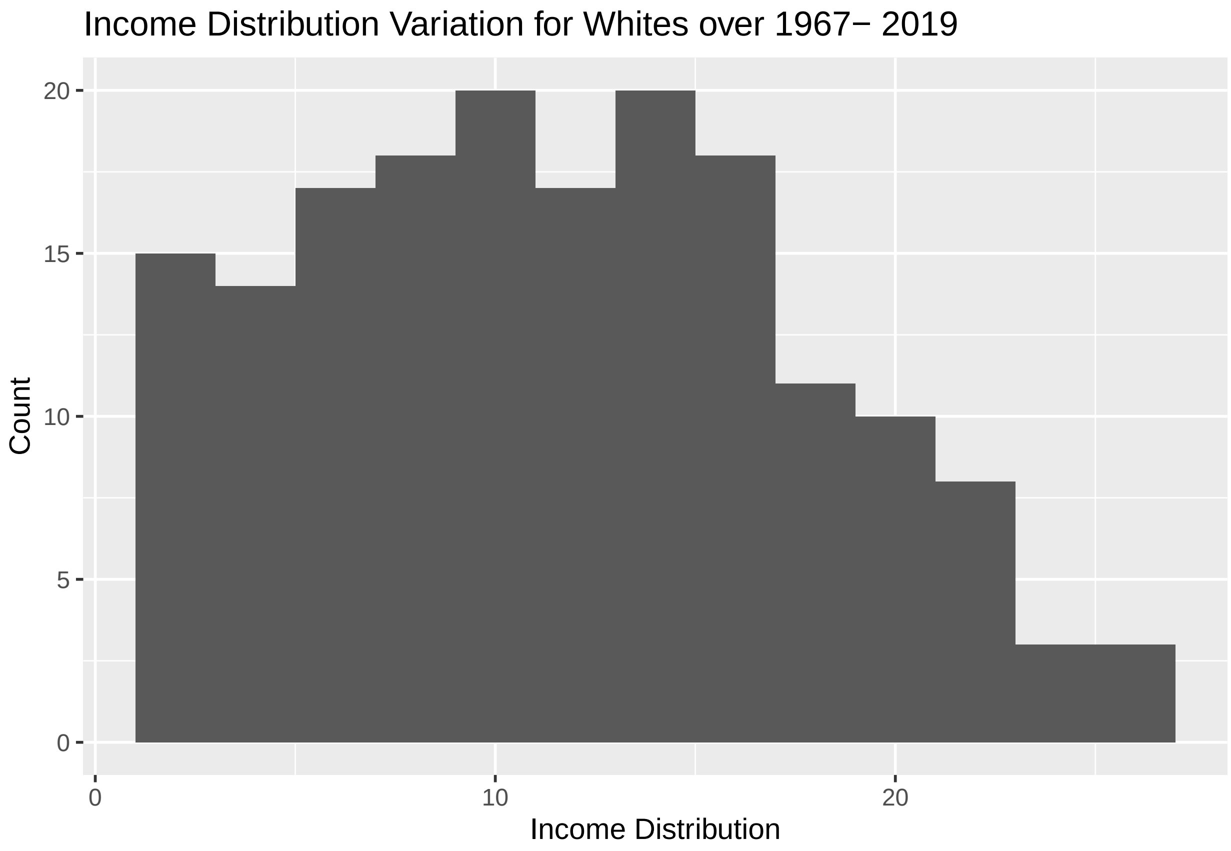
```
income_dist_3_white <- income_dist %>%
  filter(race == "White Alone",
         year >= 1967) %>%
  group_by(income_distribution) %>%
  summarize(
    n = n()
  )
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
income_dist_3_white
```

```
## # A tibble: 174 x 2
##    income_distribution     n
##                  <dbl> <int>
##  1                 1.2     1
##  2                 1.3     1
##  3                 1.5     2
##  4                 1.6     2
##  5                 1.8     2
## 6                  1.9     2
## 7                  2       4
## 8                  2.1     1
## 9                  2.2     1
## 10                 2.3     3
## # ... with 164 more rows
```

```
ggplot(income_dist_3_white, aes(income_distribution)) +
  geom_histogram(binwidth = 2) + xlab("Income Distribution") + ylab("Count") + ggtitle("Income Distribu
```



Discussion: *Depicted in the first figure above is the time series scatter plot of median household income over the time period of 1987-2019 for each of the four distinct racial groups. In comparing the qualitative trends across each of the four racial groups, the linear best fit lines suggest that the median household income has increased over time for each category, with White and Black American households on average, experiencing the most gradual climb in income, while Hispanic households appear to have achieved a slightly more pronounced increase in household income, followed by Asian households who have experienced the largest increase in median household income as evidenced by a linear trend line with the largest slope. As expected, the median household income of Blacks is the lowest followed by Hispanics, Whites, and Asians respectively, as those groups have historically experienced educational and economical inequality through discrimination and under-representation in descending order. It is interesting to note that within each racial group, there are what appears to be fluctuations in median household income that can be described by power relationships with time whose fluctuations increase in amplitude with time. Thus, while there are noticeable fluctuations, since the amplitude increases with time, a linear trend line with a positive slope is an acceptable best fit line to model the data.Now, in connecting these observations to societal and economical trends, as the value of the American dollar has changed, and the wealth gap widened over time, the median household income in the nation has also increased. This societal observation is graphically confirmed according to the aforementioned trends shown in the scatter plot. Additionally, the minor fluctuations in value for each of the groups can be* *tied to seasonal variations in the economy and market trends, and it can be observed that the fluctuations increase and decrease for each of the groups over roughly the same time period.*

Shifting gears to observing the pie charts, it is shown in the first plot that approximately half of Asian households had a household income of $100,000 or greater in the year 2019, with the proportions in the categories of "100,000 - 149,999", "150,000 - 200,000" and "200,000 and over" being roughly evenly split. This is reasonable considering that Asians have historically experienced large economic success in America. Looking further at the graph, it is shown that the proportion of Asian households with under 15,000 marginally exceeds that of the proportion of Asian households with between 15,000 and 24,999, and between 25,000 and 34,999. Shifting gears to the second plot, it is observed that approximately half of the Black households had a household income of 74,999 or less, a stark difference in trend from the Asian household distribution. To expand, approximately 17% of all Black households fell in the income bracket of 15,000 or less, roughly three times the percentage of Asian households that fell in this income bracket. This trend is reasonable considering that Blacks have historically experienced the greatest amount of economic inequality.

Looking now at the third plot, it is shown that household income is much more evenly distributed across the income brackets for white americans than for either asian or black americans with the highest frequency categories being "100,000 - 149,000" and "50,000 - 74,999". The percentage of white households with under 15,000 income lies between the black and asian groups. Lastly, looking at the pie chart depicting the hispanic income distribution, it can be observed that the disparity across income brackets is the second most pronounced of the four racial groups after that of blacks with the percentage of blacks making under 15,000 lying at roughly 10 %. It is also shown that the income bracket with the highest relative frequency among hispanics is "50,000 - 74,999". Overall, Black Americans have the greatest disparity in household income in 2019, followed by Hispanics, Whites, and Asians respectively. A significant number of blacks and hispanics made an appreciably low household income of 15,000 or under in comparison to whites and asians during this same time period, while asians experienced the greatest household income earnings.

In examining the histograms generated to answer the third question in part 1 of the project, it can be observed that the variance of income distribution is skewed at the upper-end with the highest frequency bins corresponding to income distribution ranges of (7.25, 8.75) and (17.25, 18.75) and containing a frequency of roughly 19. It is likely that the highest frequency bins correspond to the income brackets with two of the highest household income values. Shifting gears to the second histogram, it is shown that there is much more variation in the income distribution for Black Americans than for Asian Americans.In particular, the data appears to be multimodal with a mode at (1,3) and (13,17). This increase in variation suggests that there is a greater amount of disparity in income distribution for Blacks than Asians which is reasonable given the disproportionate amount of economic success that the former racial group has historically experienced relative to the latter. Looking at the distribution for Hispanics, it is observed that the shape of the data is nearly consistent with a Normal Distribution while there appears to be a mode over the values (1.25, 5). Additionally, it is observed that the variation in the distribution is more pronounced than that of Asians, but less than that of Blacks as expected. Lastly, in looking at the income distribution variation for Whites, it is observed that the data is skewed at its upper end with the highest frequency bins lying at the middle of the distribution. Additionally, this distribution is comparable in shape to that of the Asian income distribution as expected from historical trends.

Part 2

Question: *Of the following quantitative variables, "year", "income_median", "income_med_moe", and "income_distribution", which are the most and least strongly correlated/associated across each of the four primary racial groups? How is the variance in the dataset distributed across associations? Hint: Use dimensional reduction to explore these relationships .*

Introduction: *In Part 2 of this project, the "income_dist" data frame analyzed in Part 1 will again be used to explore relationships amongst quantitative variables. In particular, an observation of the correlation/association between four quantitative variables in the data frame "year", "income_median", "income_med_moe", and "income_distribution" will be made. This will be accomplished by using Principal Component Analysis as a* *form of dimension reduction of the data frame. This is a suitable method because PCA is designed to depict relationships across many variables in datasets with a large number of dimensions such as that of the data frame the author is working with. A second objective of this part of the project is to observe how the variance in the data is distributed across these relationships using the distribution of principal component variances.*

Approach: *For the first question of the second part of the project, the data will first be pre-processed by filtering the data to include the four major racial groups, and then the "select()" method will be used to retain only the "year", "race", income_median“,”income_med_moe“, and”income_distribution" data values. Then, the "unique()" method will be used to eliminate all duplicate values in the resultant frame. From here, dimensional reduction will be performed by first selecting only numeric values, and then scaling the values to a zero mean and unit variance. Then, "prcomp()" will be used to perform PCA on the resultant data frame to obtain summary statistics that quantify association amongst the different principal components in the data. Then, a scatter plot depicting the fit of the data to the first two principal components (which account for the largest variation in the data set) will be generated using "augment()", "ggplot()", and "geom_point()". Then, a rotation matrix will be produced that depicts the association between the four variables in relation to the two principal components that account for the largest variance in the data.*

For the second question of the second part of the project, the PCA fit data frame containing the summary statistics resulting from the PCA will be used to generate a relative frequency plot that depicts the distribution of the variance in the total dataset. In particular, this relative frequency plot will plot the variance explained by each principal component and the relative frequencies of the bars will quantify the percentage of total variance in the dataset captured by the associated principal component. A relative frequency plot is a suitable visualization tool for observing variance across principal components in a dataset because it allows for the user to display the proportion of values observed across qualitative/categorical variable values (such as the principal component number).

Analysis:

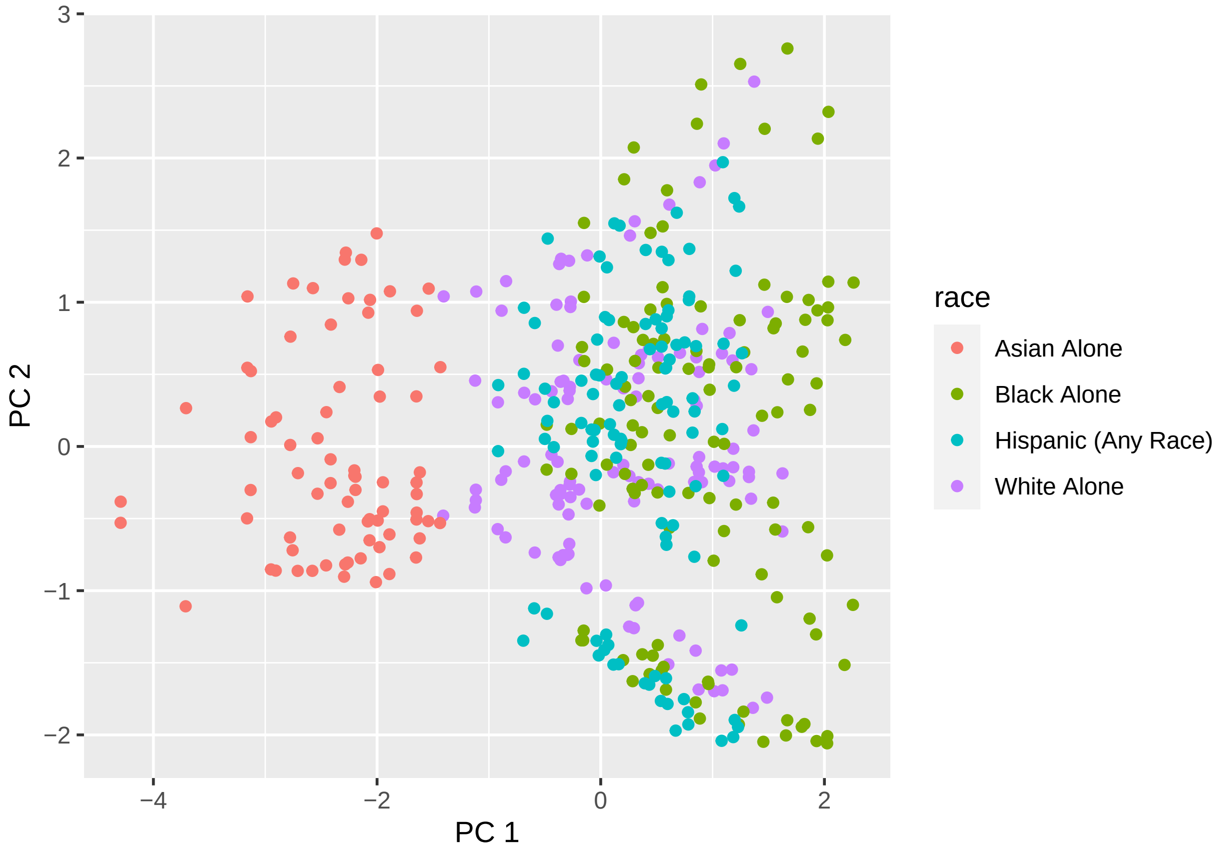
```r
# Filter data frame to include median household income values counts across four races (Asian,
#Black, Hispanic, and White)

income_dist_4 <- income_dist %>%
  filter(race == c("Asian Alone", "Black Alone", "Hispanic (Any Race)", "White Alone")) %>%
  select(year, race,income_median, income_med_moe, income_distribution) %>%
  unique() # Remove duplicate income_median values for each year and race

income_dist_4 # Output resultant data frame to console
```

```
## # A tibble: 418 x 5
##     year race        income_median income_med_moe income_distribution
##    <dbl> <chr>               <dbl>          <dbl>               <dbl>
##  1  2019 White Alone         72204            800                 8
##  2  2019 White Alone         72204            800                16.3
##  3  2018 White Alone         68156            657                 8.3
##  4  2018 White Alone         68156            657                13.1
##  5  2017 White Alone         68076            714                 8.6
##  6  2017 White Alone         68076            714                16.5
##  7  2017 White Alone         68076            714                 9
##  8  2016 White Alone         65901            585                12.3
##  9  2016 White Alone         65901            585                 7.5
## 10  2015 White Alone         64864            676                 9.4
## # ... with 408 more rows
```

```r
# Dimension Reduction and PCA

pca_fit <- income_dist_4 %>%
  select(where(is.numeric)) %>% # Select only columns with numeric (integer, double) values
  scale() %>% # Scale values to zero unit variance and mean
  prcomp() # Perform PCA

pca_fit # Output summary statistics of PCA
```

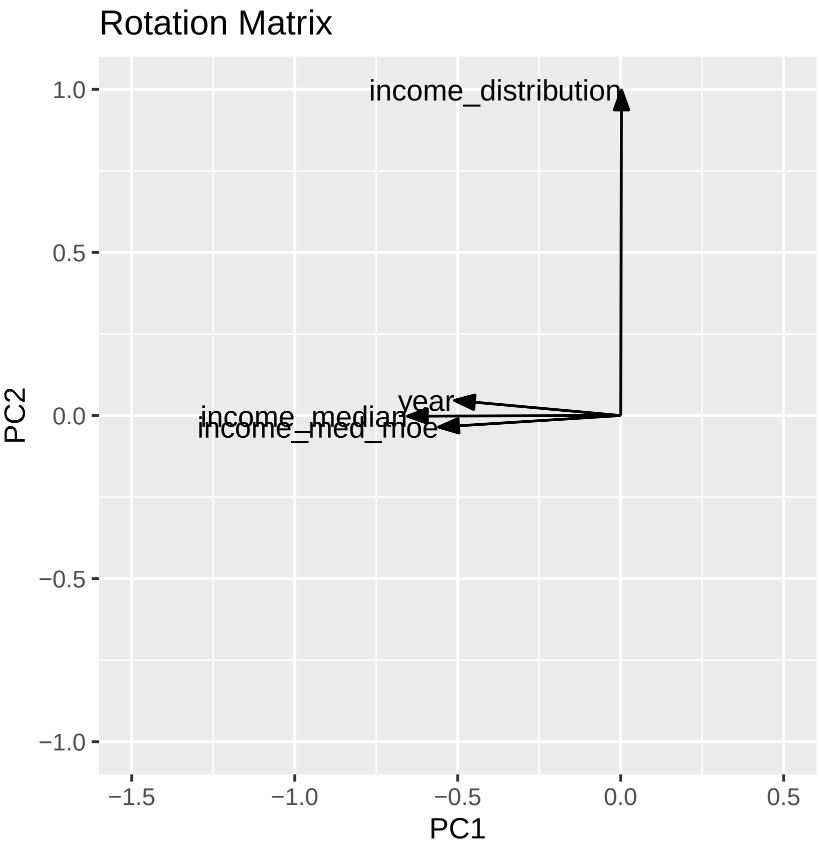
```
## Standard deviations (1, .., p=4):
## [1] 1.3189986 1.0003135 0.9026874 0.6669116
##
## Rotation (n x k) = (4 x 4):
##                              PC1          PC2         PC3           PC4
## year                -0.509940406  0.046456226 -0.76624556  0.3881627846
## income_median       -0.654176758 -0.001865586  0.05304380 -0.7544770671
## income_med_moe      -0.558575074 -0.035352135  0.63769165  0.5292385781
## income_distribution  0.002727318  0.998292825  0.05833919 -0.0007316551
```

```
library(broom)

# Perform PCA Fit

pca_fit %>%
  augment(income_dist_4) %>%
  ggplot(aes(.fittedPC1, .fittedPC2)) +
  geom_point(aes(color = race)) +
  xlab("PC 1") + ylab("PC 2")
```



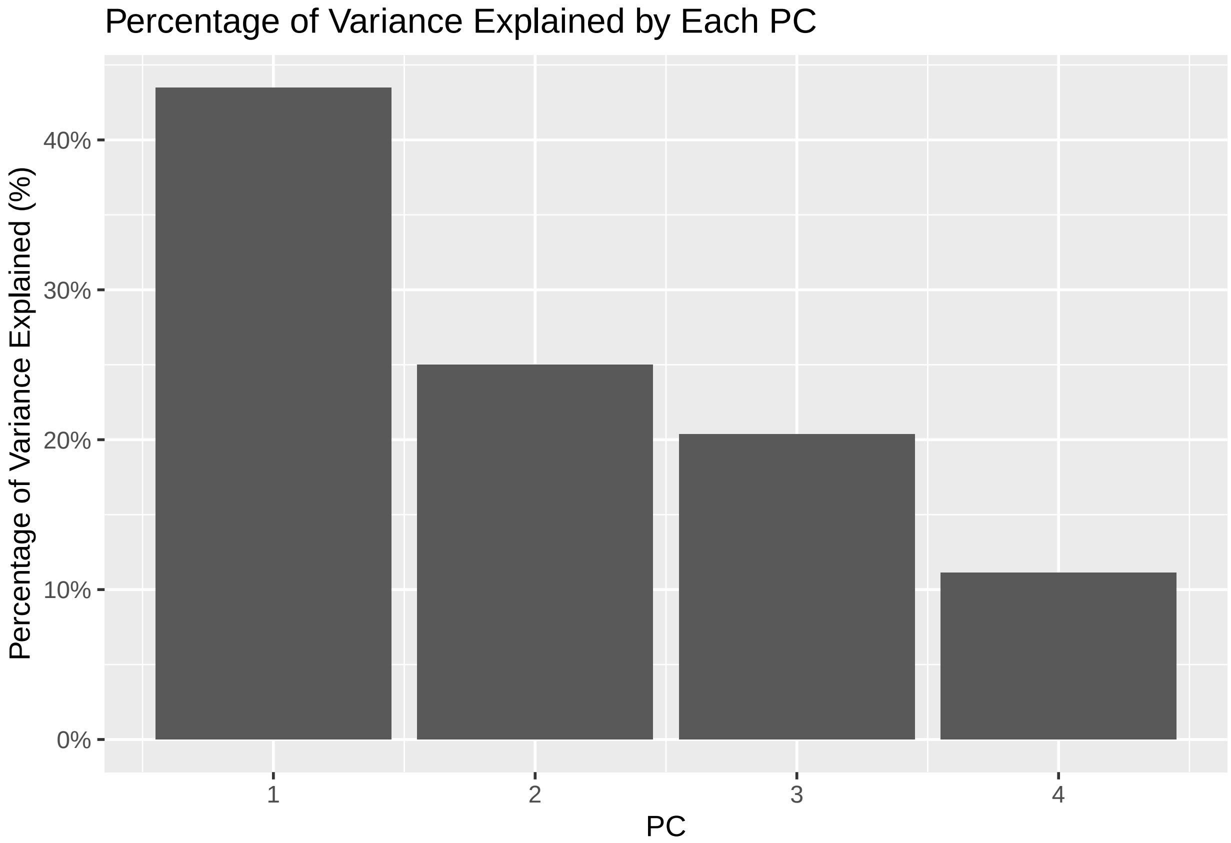
```
# Define properties of rotation matrix arrow vectors
```

```r
arrow_style <- arrow(
angle =20, length = grid::unit(8,"pt"),
ends ="first", type ="closed"
)

# Plot the rotation matrix depicting relationship between PC components 1 and 2

pca_fit %>%
# extract rotation matrix
tidy(matrix ="rotation") %>%
pivot_wider(names_from ="PC", values_from ="value",
names_prefix ="PC"
) %>%
ggplot(aes(PC1, PC2)) +
geom_segment(
xend =0, yend =0,
arrow = arrow_style
) +
geom_text(aes(label = column), hjust =1) +
xlim(-1.5,0.5) + ylim(-1,1) +
coord_fixed() + ggtitle("Rotation Matrix")
```



```r
# Visualizing Variance across PC Categories

pca_fit %>%
# extract eigenvalues
```

```
tidy(matrix ="eigenvalues") %>%
ggplot(aes(PC, percent)) +
geom_col() +
scale_x_continuous(
# create one axis tick per PC
breaks =1:4
) +
scale_y_continuous(
name ="Percentage of Variance Explained (%)",# format y axis ticks as percent values
label = scales::label_percent(accuracy =1)
) +
  ggtitle("Percentage of Variance Explained by Each PC")
```



Discussion: *In examining the scatter plot of the PCA analysis of the entire dataset containing the "year", "income_median", "income_med_moe", and "income_distribution", it is shown that the cluster for the Asian households is distinct compared to the clusters of the Hispanic, White, and Black households which contain a large amount of overlap. Because of this trend, it can be concluded that Asian household data are separated from the Hispanic, White, and Black household data along the first principal component. This suggests that there is a lot of disimilarity in the Asian household data from the collective Hispanic, White, and Black household data across the first principal component. In examining the second cluster, it is observed that the Hispanic and White household data does not separate much over principal components 1 or 2, but that the Black household data separates noticeably from both along principal components 1 and 2. A takeaway from these trends is that Black household income data has a small-moderately larger disparity than its White and Hispanic counterparts, and that Asian household income data has a much smaller disparity than White, Hispanic, and Black data collectively.*

In examining the rotation matrix in the second figure, it is shown that the income median, income median margin of error, and year are all very strongly correlated. In particular, they have a strong positive correlation with a near perfect proportional relationship as evidenced by the close relative spacing of the vectors. The sense of these vectors suggests that all variables contribute negatively to the first principal component, which represents the overall household income data. An interpretation of this relationship is that as time has increased from 1967-2019, the median household income, and the median household income margin of error across all racial groups has increased through a nearly direct relationship. This is reasonable given that it is common knowledge that median household incomes in the nation have increased with time as the price of the American dollar, and the economy have changed. It is also reasonable that the median margin of error would increase over this time as the margin of error of a variable is typically directly proportional to the value of the variable itself, hence, since the median income has increased, so too would the median income margin of error. An interesting observation that can be made from the figure is that the income distribution appears to have no correlation/association with the median income, median income margin of error, or the year as evidenced by a vector that is nearly orthogonal to the other three. This can be contextualized by understanding that disparities across income distributions have not changed over time, nor with a change in household income namely, the wealth gap amongst the rich and the poor has changed very little. Additionally, the second principal component represents the difference between the income median margin of error and the income distribution.

In examining the relative frequency plot shown in the last figure, it can be observed that the first principal component accounts for nearly half of the total variance in the dataset at approximately 43%, while components 2-4 account for approximately 25%, 20%, and 12% of the total variance respectively. A takeaway from this trend is that the overall household income values account for approximately 43% of the variation in the data's various measurements.