ALY 6000
Intro To Analytics
Descriptive Statistics and visualizations
"Assignment-4"

PROFESSOR – ROY WADA

Submitted By

ITI ROHILLA

On

Tue Oct 17

OVERVIEW

This dataset is about substance abuse (cigarettes, marijuana, cocaine, alcohol) among different age groups and states. Data was collected from individual states as part of the NSDUH study.

The data ranges from 2002 to 2018. Both totals (in thousands of people) and rates (as a percentage of the population) are given.

Below is the Data Dictionary for the data set used:

1. **State (String):**
   Description: The state that this report was created for.
   Example Value: "Alabama"
2. **Year (Integer):**
   Description: The year that this report was created for.
   Example Value: 2002
3. **Population.12-17 (Integer):**
   Description: Estimated population for this age group (12 to 17 year olds) in this year from US Census data for this state.
   Example Value: 380805
4. **Population.18-25 (Integer):**
   Description: Estimated population for this age group (18 to 25 year olds) in this year from US Census data for this state.
   Example Value: 499453
5. **Population.26+ (Integer):**
   Description: Estimated population for this age group (26 years old or older) in this year from US Census data for this state.
   Example Value: 2812905
6. **Totals.Alcohol.Use Disorder Past Year.12-17 (Integer):**
   Description: The estimated number of people (in thousands) that have a use disorder on alcohol in the past year among this age group.
   Example Value: 18
7. **Totals.Alcohol.Use Disorder Past Year.18-25 (Integer):**
   Description: The estimated number of people (in thousands) that have a use disorder on alcohol in the past year among this age group.
   Example Value: 68
8. **Totals.Alcohol.Use Disorder Past Year.26+ (Integer):**
   Description: The estimated number of people (in thousands) that have a use disorder on alcohol in the past year among this age group.
   Example Value: 138
9. **Rates.Alcohol.Use Disorder Past Year.12-17 (Float):**
   Description: Amount per 1000 people that have a use disorder on alcohol in the past year among this age group.
   Example Value: 48.336
10. **Rates.Alcohol.Use Disorder Past Year.18-25 (Float):**
    Description: Amount per 1000 people that have a use disorder on alcohol in the past year among this age group.
    Example Value: 136.49
11. **Rates.Alcohol.Use Disorder Past Year.26+ (Float):**
    Description: Amount per 1000 people that have a use disorder on alcohol in the past year among this age group.
    Example Value: 49.068
12. **Totals.Alcohol.Use Past Month.12-17 (Integer):**
    Description: The estimated number of people (in thousands) that have used alcohol in the past month among this age group.
    Example Value: 57
13. **Totals.Alcohol.Use Past Month.18-25 (Integer):**
    Description: The estimated number of people (in thousands) that have used alcohol in the past month among this age group.
    Example Value: 254

NOTE- Remaining all the columns of the data set are Totals and Rates for different substance, Age, user type which will follow the same pattern as specified in the above Data Dictionary.

Visualization-1

Q1. Calculate the average rate of alcohol use disorder for each age group and all states for the years 2002, 2003, 2004, and 2005.Remove any rows that contain NAs.Create a multi-bar chart for the following visualization.

A1.

Filter the data to select only the rows for the specified years (2002, 2003, 2004, and 2005).

Selects the relevant columns for alcohol use disorder rates for the three age groups (12-17, 18-25, 26+)

Groups the data by year and calculates the average rate for each age group using the summarise function.

The resulting data frame, ***avg_alcohol_rates***, contains the average rates for each age group for the specified years.

```
> highchart() %>%
+   hc_chart(type = 'multiBarChart') %>%
+   hc_xAxis(categories = avg_alcohol_rates$Year) %>%
+   hc_add_series(name = 'Avg_Rate_12_17', data = avg_alcohol_rates$Avg_Rate_12_17, type = 'column') %>
+   hc_add_series(name = 'Avg_Rate_18_25', data = avg_alcohol_rates$Avg_Rate_18_25, type = 'column') %>
+   hc_add_series(name = 'Avg_Rate_26_plus', data = avg_alcohol_rates$Avg_Rate_26_plus, type = 'column')
+   hc_title(text = "Multi-Bar Bar Chart") %>%
+   hc_subtitle(text = "Comparison of Multiple Age_Groups by Years") %>%
+   hc_yAxis(title = list(text = "Avg_Rate_Of_Alcohol_Intake_PastYear")) %>%
+   hc_legend(enabled = TRUE)
```

NOTE:

- In this example, we use '**highcharter**' to create an interactive multi-bar bar chart with three variables (Avg_Rate_12_17, Avg_Rate_18_25, Avg_Rate_26_plus) for each category.
- The type = '**multiBarChart**' specifies the type of chart. For this we need to **install** and **load** the '**highcharter**' package.
- We can interact with the chart, zoom in, pan, or get more information about the data points.

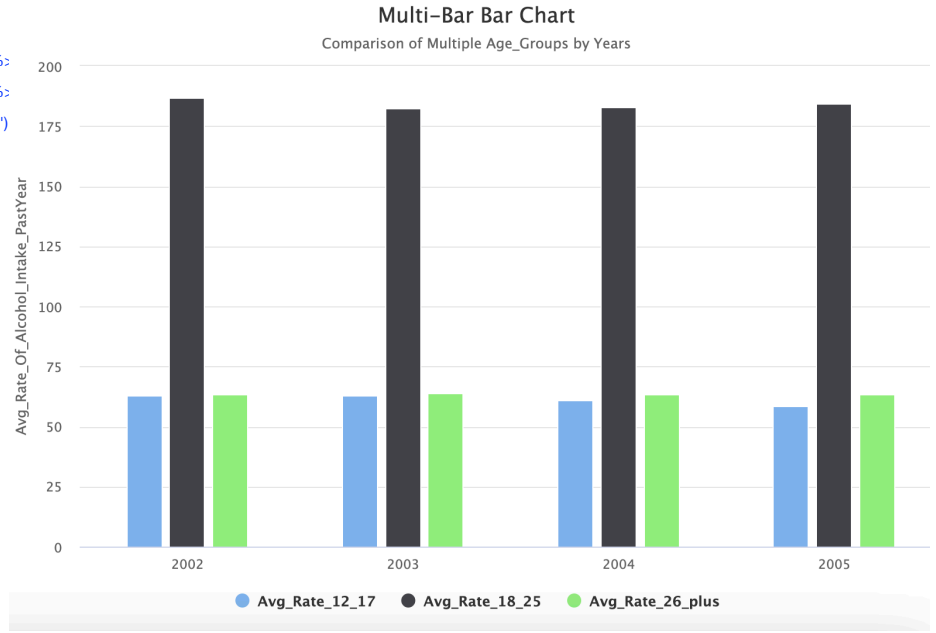


Fig 1: Avg_Alcohol_Rates of each age group for specific Years

Visualization-2

Q2. Display the New Users of Marijuana in thousands between the age of 12-17 . Create an interactive geo - spatial graph using plotly library of R showing all the states of America.

A2.

For creating geo-spatial graph install and load the library(plotly).

Store the data of .csv file that contains abbreviations(code) of names of US in a vector and join it using inner join with the data frame we create.Codes will be used when mouse pointer hovers over the states to see information.

Create a data frame by selecting Year, State, Code and Totals.Marijuana.New Users.12-17.

```
> Marijuana_graph <- plot_geo(Marijuana_df,
+                     locationmode = 'USA-states',
+                     frame = ~Year) %>%
+  add_trace(locations = ~Code,
+          z = ~y1,
+          zmin = 0,
+          zmax = max(Marijuana_df$y1),
+          color = ~y1,
+          colorscale = "Earth",
+          colorbar = list(title = "Totals_Age=12-17")) %>%
+  layout(geo = list(scope = 'usa'),
+        title = '\n New Users of Marijuana in the US\n 2002-2018',
+        annotations = list(
+          list(
+            x = 0.5,
+            y = -0.1,
+            showarrow = FALSE,
+            text = "Note: This chart displays the total number of new marijuana users aged 12-17 in totals
+            xref = "paper",
+            yref = "paper"
+          )
+        )
+ )
```

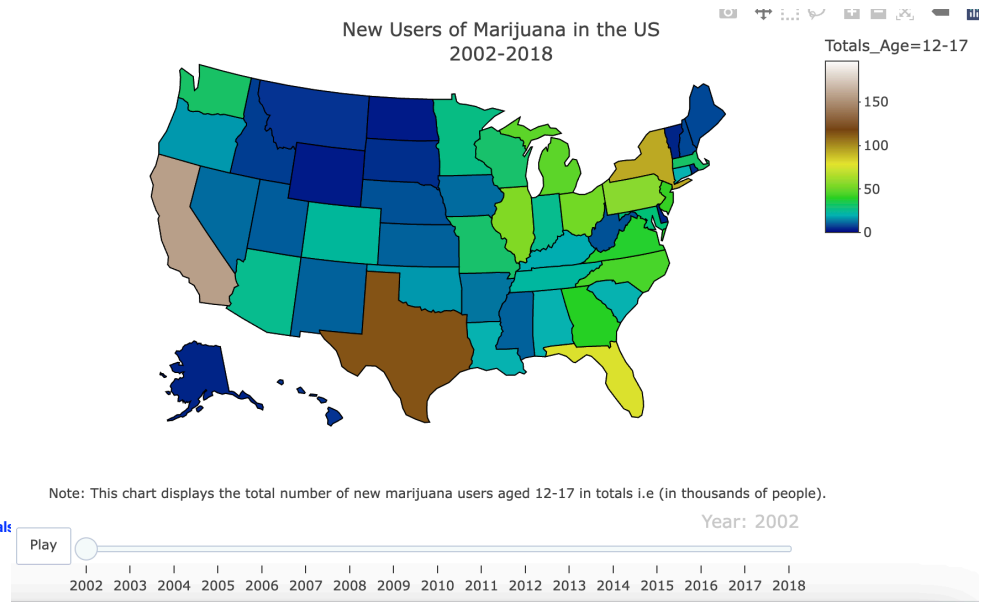


Fig 2: Graph Depicting the New Users of Marijuana in US between the age 12-17

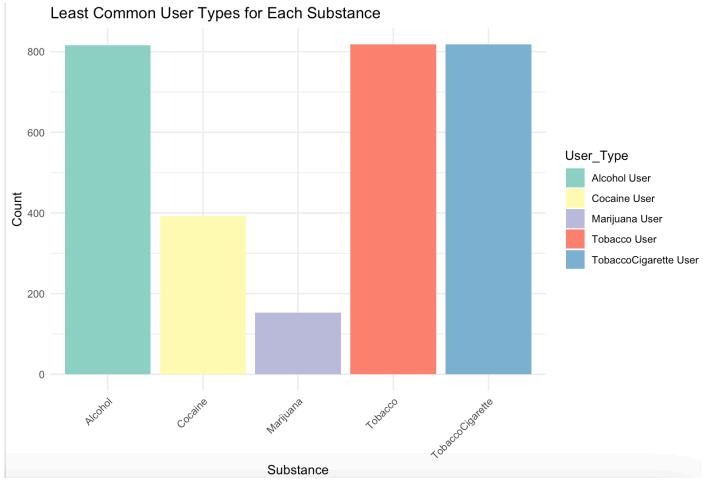
Visualization-3

Q3.What are the most and least common user types for each substance group? Display using a R chart to show Visualization.

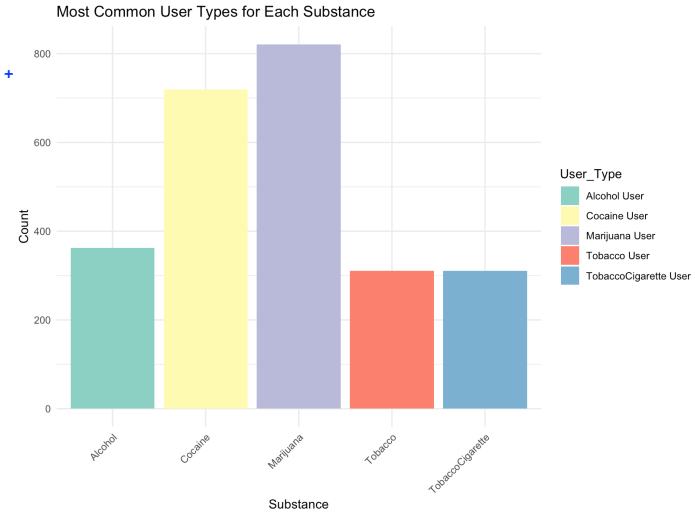
A3.

- Calculate the total number of users for each substance within the 18-25 age group by adding users of each substance type.
- Create a new dataframe to store the total users for each substance.
- Find most and least common user types for each substance using min and max function
- Create new data frames for most_common_df and least_common_df
- Create bar graphs for each using ggplot.

```
> ggplot(most_common_df, aes(x = Substance, y = Count, fill = User_Type)) +
+   geom_bar(stat = "identity") +
+   labs(title = "Most Common User Types for Each Substance",
+        x = "Substance",
+        y = "Count") +
+   scale_fill_brewer(palette = "Set3") +
+   theme_minimal() +
+   theme(axis.text.x = element_text(angle = 45, hjust = 1))
```




```
> ggplot(least_common_df, aes(x = Substance, y = Count, fill = User_Type)) +
+   geom_bar(stat = "identity") +
+   labs(title = "Least Common User Types for Each Substance",
+        x = "Substance",
+        y = "Count") +
+   scale_fill_brewer(palette = "Set3") +
+   theme_minimal() +
+   theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



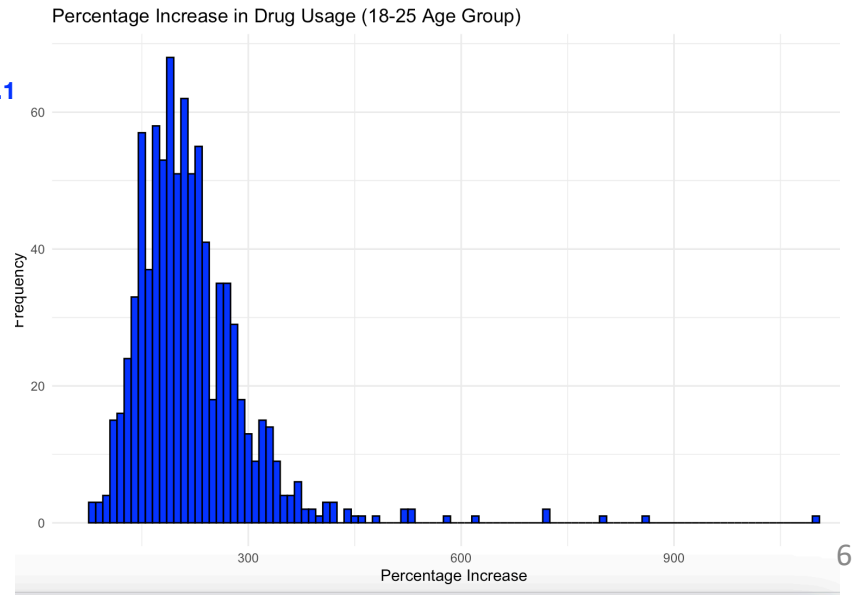
Visualization-4

Q4.What are Calculate the percentage increase in drug usage over the years for the 18-25 age group compared to the 12-17 age group?.Create a histogram to visualize the percentage increase.

A4.

- Load the libraries and calculate percentage increase in drug usage.It does this by subtracting the drug usage for ages 18-25 in the past year (Totals.Marijuana.Used Past Year.18-25) from the drug usage for ages 12-17 in the past year (Totals.Marijuana.Used Past Year.12-17). The result is divided by the drug usage for ages 12-17, and then it's multiplied by 100 to convert it into a percentage.
- Histogram that visualizes the distribution of the percentage increase in drug usage for the 18-25 age group. It provides insights into how drug usage has changed over time within this specific age groupFind most and least common user types for each substance using min and max function
- The histogram's bars represent different levels of percentage increase, and the frequency of each level is shown on the y-axis.

```
> data <- data %>%
+   mutate(PercentageIncrease1825 = ((`Totals.Marijuana.Used Past Year.1
Past Year.12-17`) * 100)
> ggplot(data = data, aes(x = PercentageIncrease1825)) +
+   geom_histogram(binwidth = 10, fill = "blue", color = "black") +
+   labs(
+     title = "Percentage Increase in Drug Usage (18-25 Age Group)",
+     x = "Percentage Increase",
+     y = "Frequency"
+   ) +
+   theme_minimal()
```



OBSERVATION & FOLLOW-UP

1. Multi-Bar Bar Chart:

- For creating Multi-Bar Bar Chart we used **Highcharter** , it is important because it empowers R users to create stunning, interactive data visualizations and enables them to effectively communicate insights from their data to a broader audience.
- It bridges the gap between data analysis in R and creating engaging web-based data visualizations without requiring extensive web development skills.

FOLLOW-UP : Are there significant changes or fluctuations in these rates for different age groups?

2. US Map Visualization:

- This utilizes **Plotly** to create a U.S. map showing the total number of new marijuana users aged 12-17 across different states from 2002 to 2018.
- Geospatial visualization with tools like Plotly is essential for unlocking valuable insights from geographic data. It allows data analysts, researchers, and decision-makers to harness the power of location-based data and effectively communicate findings to a broad audience.
- The combination of geospatial visualization and Plotly empowers users to explore, analyze, and present geographic data with clarity and interactivity.

FOLLOW-UP :Create a US Map Visualization that shows all the data of the file in one single visualization using interactive legends .

3. Most and Least Common User Types:

- The code calculates the most and least common user types for different substances (Alcohol, Tobacco, Cocaine, Marijuana, and TobaccoCigarette) within the 18-25 age group.
- This demonstrates the creation of bar graphs using **ggplot2** in R. It emphasizes the importance of data preparation, aesthetics mapping, and customization options to produce clear and visually appealing bar charts for analyzing and presenting data related to the most and least common user types for different substances.

4. Histogram for Percentage Increase:

- It demonstrates the process of calculating and visualizing the percentage increase in drug usage for a specific age group using a histogram. It's a useful approach to understand the distribution and trends in drug usage within the specified age group.
- **Dplyr** package to perform data transformation. Specifically, it calculates the percentage increase in drug usage for a specific age group (18-25) over the years.

FOLLOW-UP :How has the percentage increase in drug usage for the 18-25 age group evolved over time, and are there any noticeable trends or patterns in the histogram that might suggest changes in drug usage behavior within this age range?

Overall, these reveal visualizations and analyses related to drug usage, alcohol use disorder rates, and trends in specific age groups. Further exploration could provide insights into the changing patterns and prevalence of substance use and user types.

CITATIONS

- R Maps: Beautiful Interactive Choropleth & Scatter Maps with Plotly
  https://www.youtube.com/watch?v=RrtqBYLf404

- List-of-US-States used in geo-spatial graph.
  https://github.com/jasonong/List-of-US-States/blob/master/states.csv

- R with Highcharts visualisations using Highcharter library
  https://www.youtube.com/watch?v=of8ras0Bl8Q

  https://www.geeksforgeeks.org/creating-interactive-plots-with-r-and-highcharts/

- Source of Data set chosen
  https://corgis-edu.github.io/corgis/csv/drugs/