Assignment-4

Divya, Nupur, Ranadheer, Varun

Data Description:

The dataset State-of-homelessness has data structured into 541 columns with 58 rows. It consist of 14 sheets having data over 14 years beginning from 2007 to 2020. Data captures information regarding different sections of the population in the United States experiencing homelessness between last one and a half decade. Data is organized in the following manner:

Age: Under 18; 18 - 24; Above 24
Gender: Male, Female, Transgender, Non-confirming Gender
Race: Hispanic, Non-Hispanic, White, African/Black-American, Asian, American-Indian, Native Hawaiian, Multiple Races and Other Races
Family status: In-family, Individual, Household
Other: Veterans, Unaccompanied Youth, Parenting Youth

The above-mentioned criteria are paired with categories such as

Overall homeless
Sheltered ES homeless
Sheltered TH homeless
Sheltered SH homeless
*Sheltered Total homeless and*
Unsheltered Overall

resulting in segregation of the data into 541 columns.

Data Cleaning-

The data for these columns has been added into rows on a state-by-state basis in alphabetical order with the 58th row being a summation(total) of the earlier 57 rows. We have worked on data summed yearly to consider the trend more peculiarly. It is to be noted however that the abbreviations of the states have been considered instead of their actual names. Data for every year from 2007 up until 2020 has been structured similarly and has been separated into different sheets in the excel file. The very first column of every year mentions the number of CoC's. Regions outside the mainland but are an integral part of the USA such as Alaska and Hawaii have also been taken into consideration for this purpose.
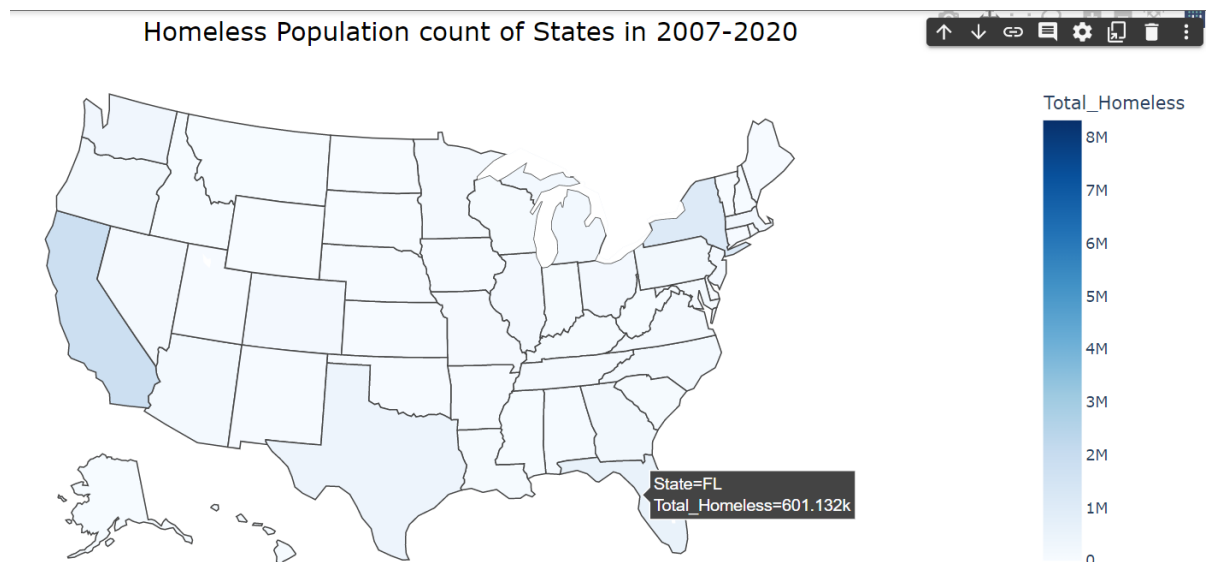
Understanding the Visualizations:

1. **WordCloud-**
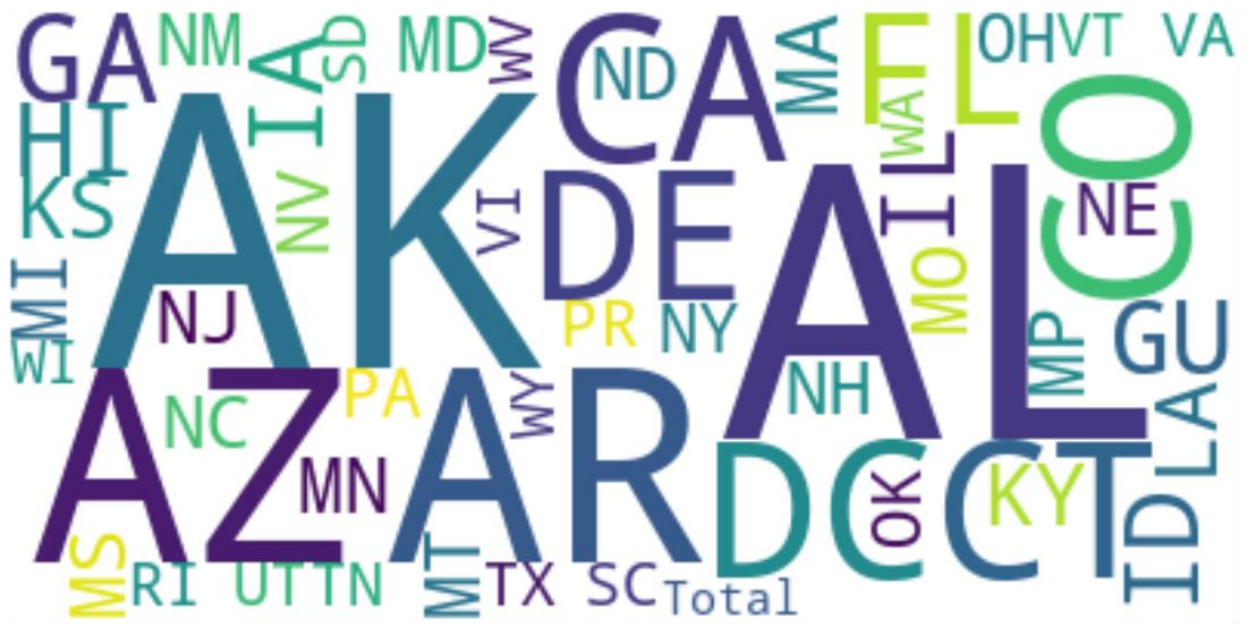
We have used WordCloud Visualization for representing the text data of States as Abbreviation on the Total Homeless Dataset for the span of 14 years (2007-2020). The size of word signifies the corresponding data points of Homeless count present in that particular state.
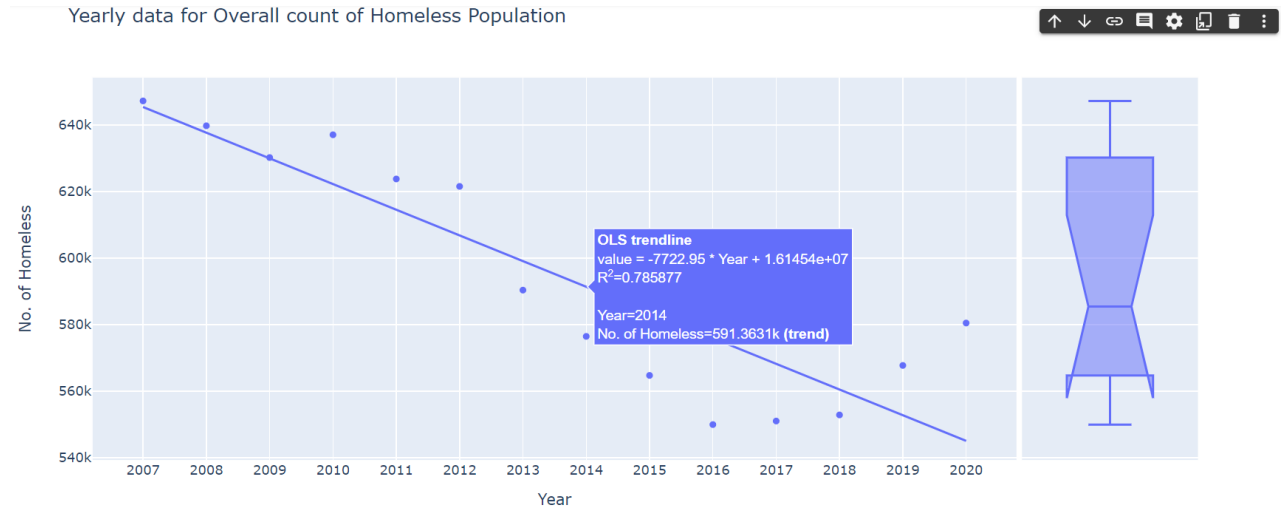
2. **Choropleth-**

The above-mentioned plot is a Choropleth which is a type of Map visualization composed of polygonal outline based on states and union territories. The Choropleth is mapped on the Total Homeless Data frame in which Total column is serving as the scale of shade. Each State depicts a shade and corresponding Total number of Homeless people present in that state in span of 14 years. California has the highest Homeless people count in last one and a half decade.

3. Stacked Bar Chart-
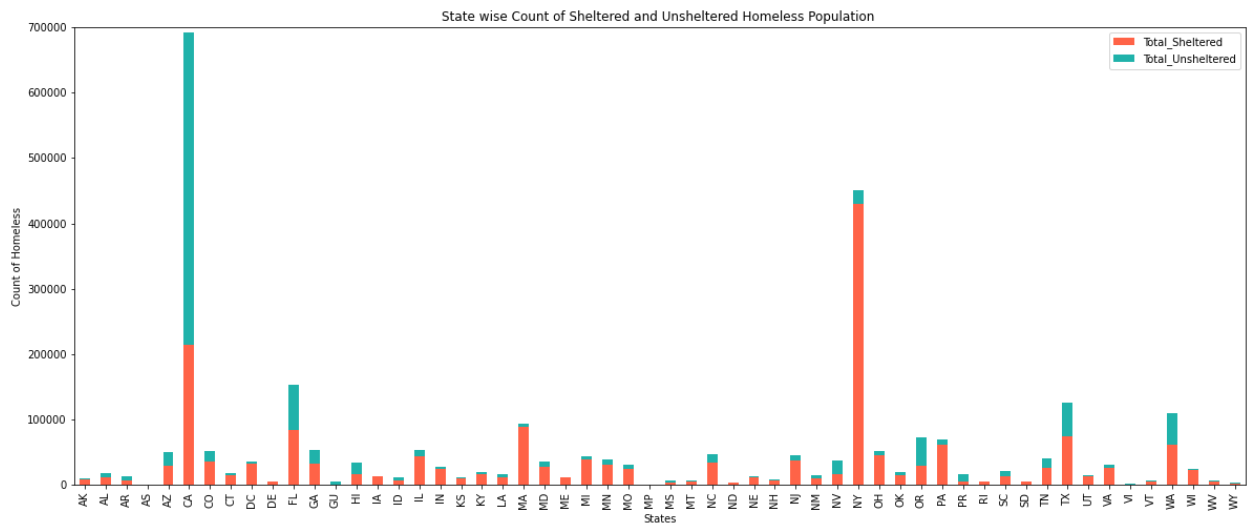
The above-mentioned Stacked Bar chart is used to visualize the difference between the Total Sheltered Homeless and Total Unsheltered Homeless People in last 5 years state wise. Summation of Sheltered and Unsheltered People from 2016-2020 is taken into the consideration which gives us following interesting insight like New York and California has the highest Total Sheltered Homeless while Florida and California again has peaked in Total Unsheltered Homeless People tally.

4. Scatter Plot (with trendline and Box Plot)-

Yearly data for Overall count of Homeless Population

OLS trendline
value = -7722.95 * Year + 1.61454e+07
R^2=0.785877

Year=2014
No. of Homeless=591.3631k (trend)

This scatter plot is based on Total Overall Homeless People counted annually from 2007-2020 irrespective of the state distribution. The slope of Trendline clearly explains the fact that the count of Overall Homeless people in USA is gradually decreasing in last one and half decade. Though there is a minimal rise in 2019-2020's Total Homeless Count but overall, the trend is still retarding on Overall Homeless Count.

5. Bar graph-

YoY Comparison of Demographic data for Homeless
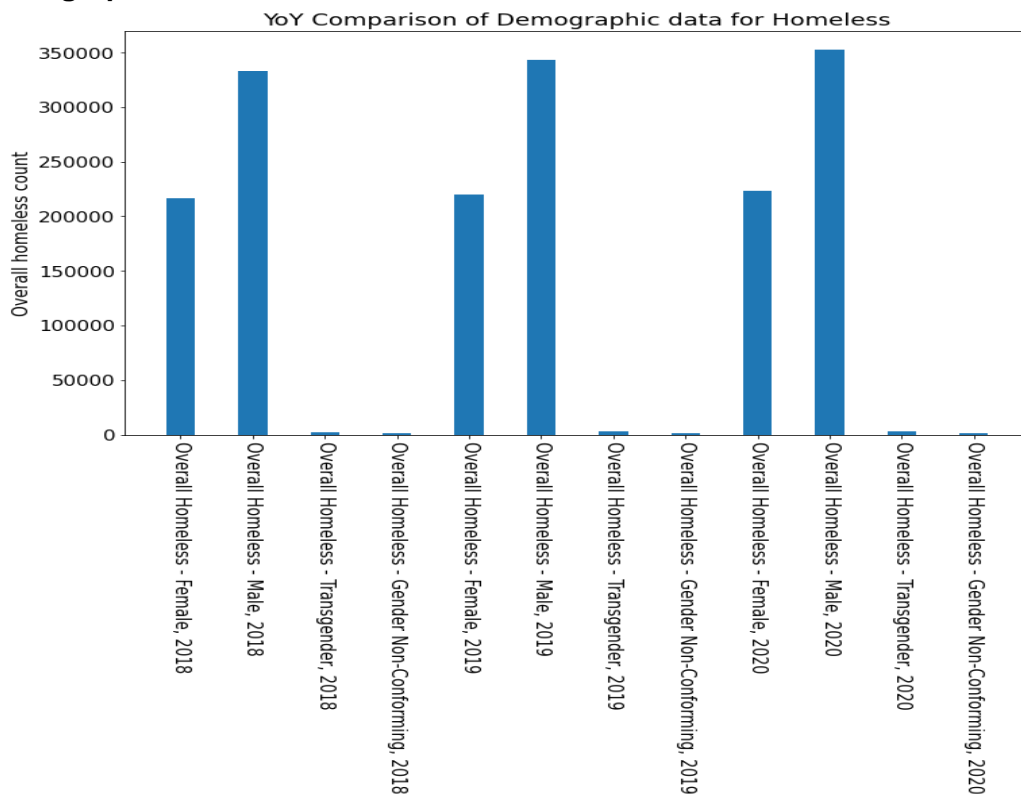
The above-mentioned bar chart is used for visualizing the Demographic distribution of Homelessness Data. The data of Male, Female, Transgender and Non-conforming Gender is under the light for 2018-2020. Clearly the count of Homeless people has ticked in with Male, Female and Transgender increasing with marginal differences.

6. **FacetGrid (Lineplot)-**
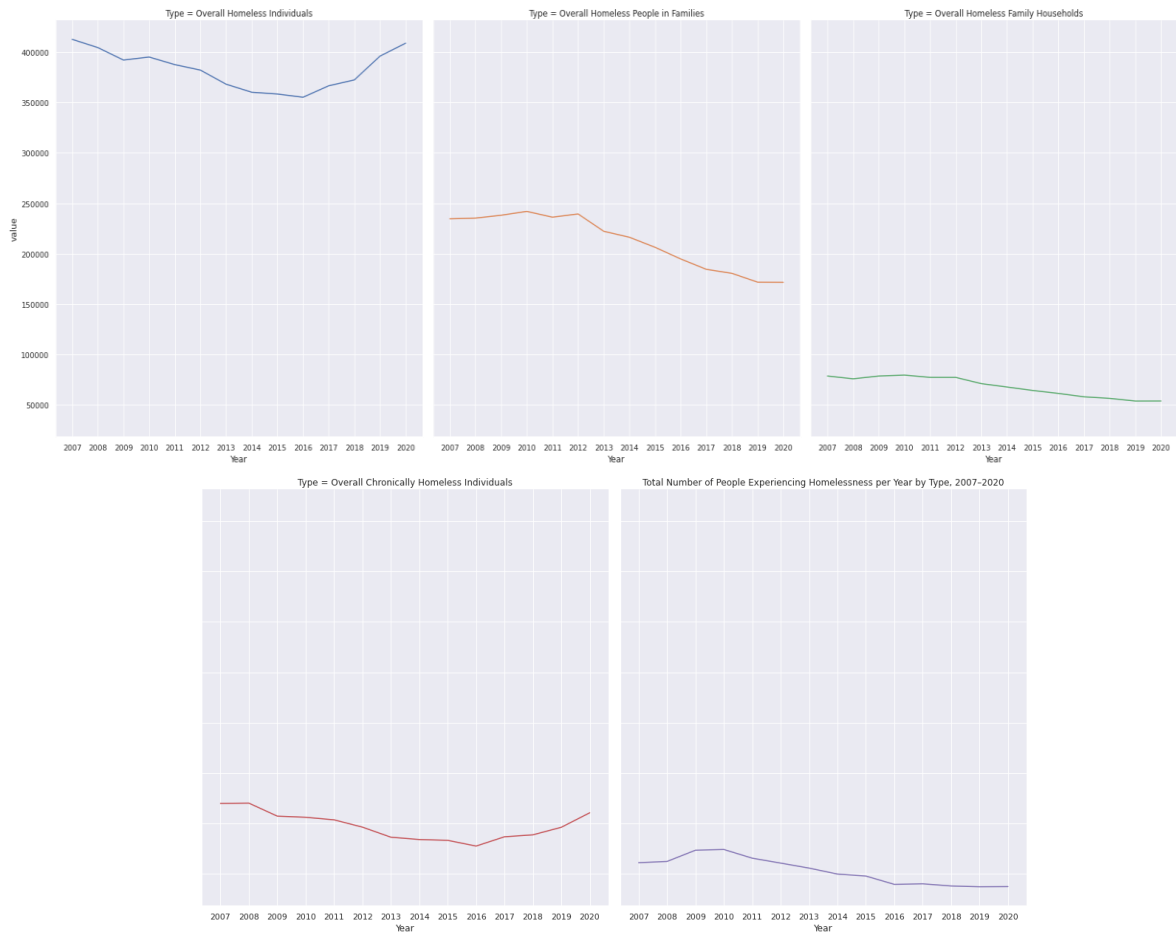
The Facetgrid applying Lineplot is plotted on the Types data frame which includes annual summation of Homeless people based on Veterans, Chronical, Families and Individual year wise. So, the Trend in Homeless people in families, Homeless Family Households and Veterans is first constant and then retarding in given span of time, while pattern in Homeless Individuals and Homeless Chronically is similar where count of Homeless people decreased for both simultaneously in first half of the data till 2014 and then it started increasing in next half till 2020 in both of them.

7. Pie Chart-
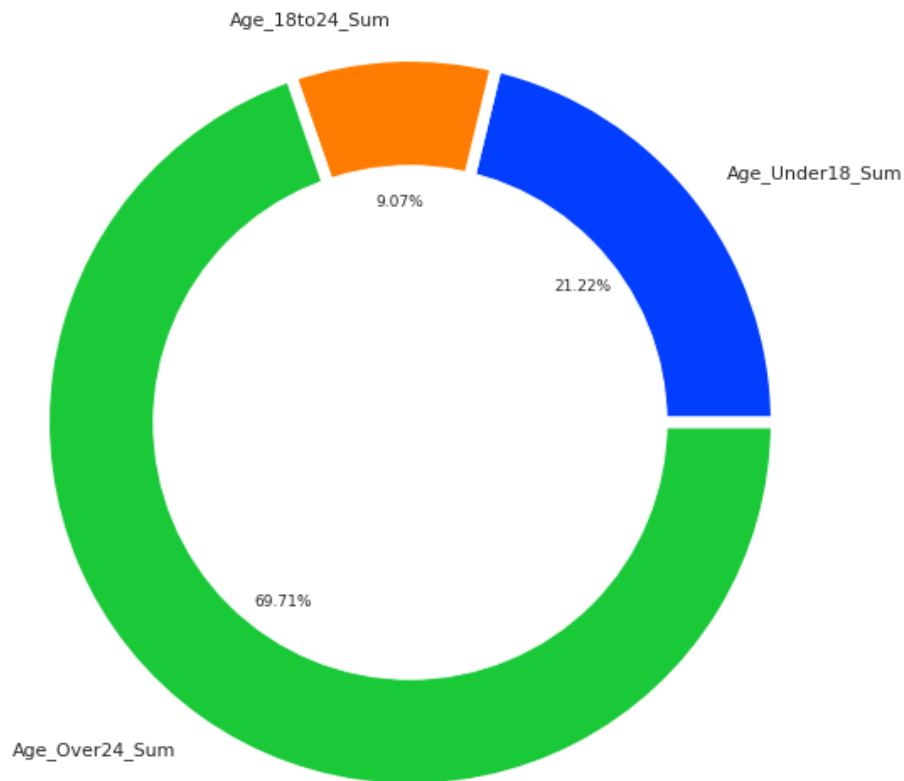
Pie Chart here is used to understand the Age based category present in data having three subsets namely: Age under 18, Age Between 18-24 and Age above 24. Considering the annual data from 2013-2020, the evident fact is people over 24 years are more prone to being Homeless, followed by the category of people under 18 years of range and finally the people in range of 18 to 24 years marks their place being least in Homeless Age category.

8. Label Lines with Annotations-
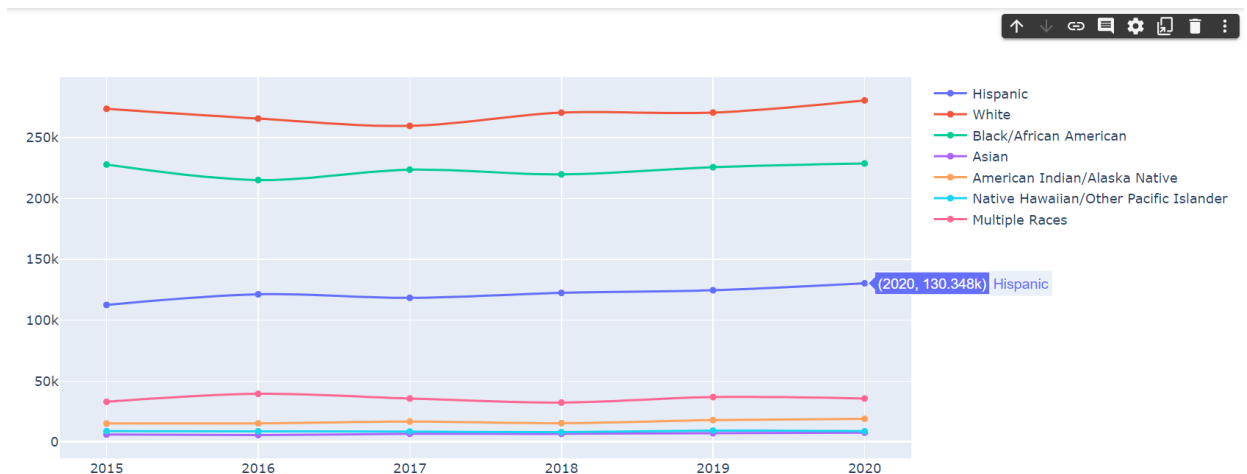
Label Line visualization remarkably signified the distribution of Homeless People based on their Races over the time period of 6 years starting from 2015-2020 accounted per annum. The Trend in American Indian/Alaska Native, Native Hawaiian/Other Pacific Islander and Asian is constant over the years, while other Multiple Races had two tides of increase and decrease in count of Homeless People in almost equal half of 4 years. Homeless people of Hispanic increased over the time, while Black/African American and White label line of races showed constant fluctuation every year.

Conclusion-

The 2007_2020_PIT_State data set was used to compare the sales attributes of various models of laptops, across 14 different sheets, with different sets of technical features. Various plots have been laid out systematically to analyze the varying trends in the Homeless count of people based on the varying parameters. Major trends were captured using the graphical tools like Scatter plot, Choropleth, Word Cloud, Bar Graph and Label Line with annotations to name a few. An important inference from various plots is that there is a constant decrease in number of Homeless people talking overall in span of 14 years in USA on a constraint of Yearly data.