Press ALT + F8 to see a list of keyboard shortcuts

This is a graded discussion: 25 points possible

due Feb 15

64 69

Flash Paper

For this assignment, you will write a "flash paper" (3-4 paragraphs) discussion post about the dataset you chose in Homework 2 Part A and discuss your exploratory analysis of the data. Your discussion post will be shared with your classmates.

Post your text directly into the discussion post. Export any figures as images (e.g., png or jpg) and include them directly in the post with the text. (You may prefer to prepare your post in a separate text editor and then paste into your browser, but please do not upload separate documents for this assignment.)

Your post should be formatted as follows:

1. Describe the dataset and where it comes from (making sure to cite the data source). Explain why you chose this dataset and what questions you wanted to explore in your visualization.

2. Describe the structure of the dataset and the variables of interest. Describe any preprocessing needs (tidying, cleaning, transformation, etc.) and describe the steps you took to perform the preprocessing.

3. Present at least 1 figure that is interesting to you and describe your observations and any key takeaways from the visualization and your exploration of the dataset.

Post your submission as a discussion post below. You will be able to view your classmates' submissions as well.

param Unread

Subscribe Subscribed

Reply

Attach

Cancel Post Reply

- 1

- 2

Ameya Santosh Gidh (*He/Him*) (https://northeastern.instructure.com/courses/170748/users/144937)

Feb 5, 2024

Title: Exploring Netflix Content Trends Over Time Dataset Description: I chose the Netflix dataset s.

Title: Exploring Netflix Content Trends Over Time

1. **Dataset Description:**

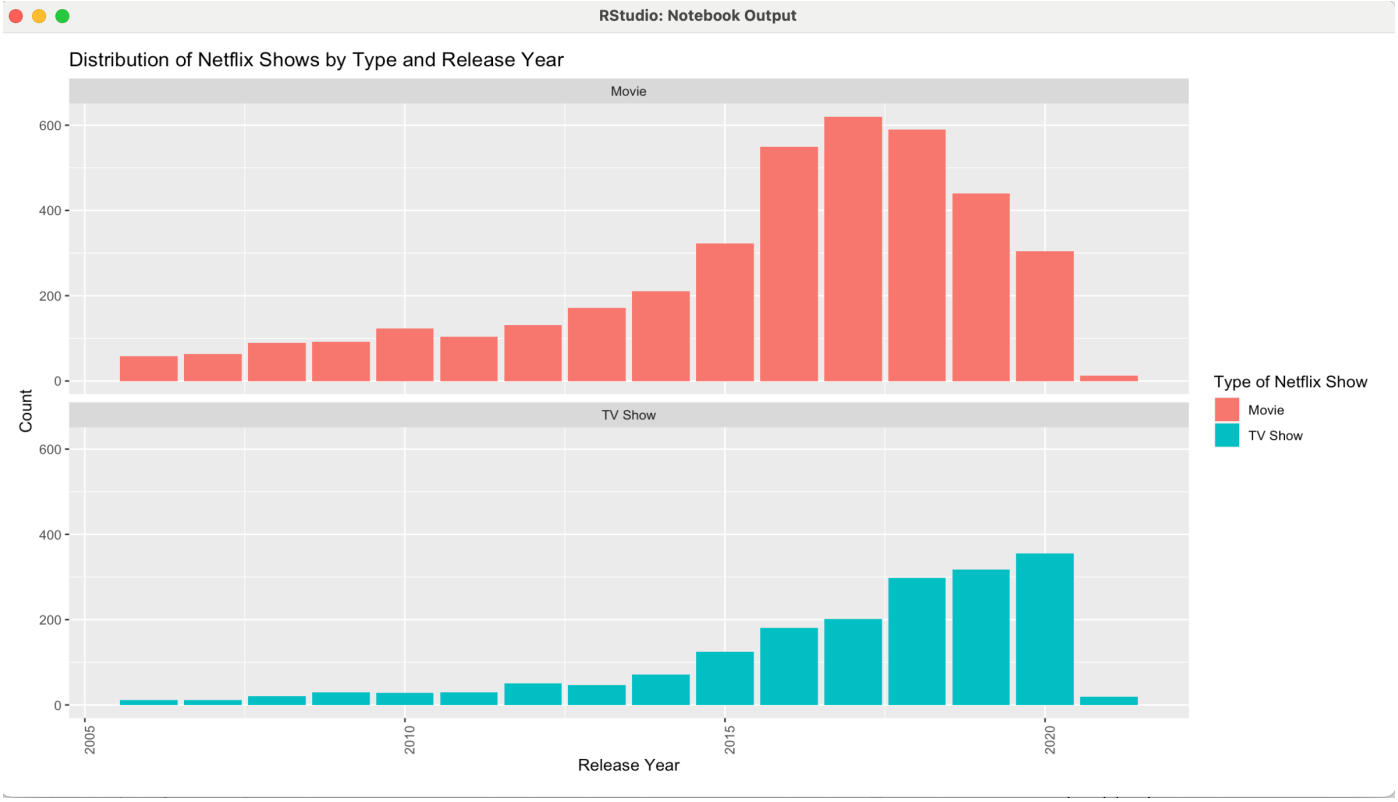
I chose the Netflix dataset sourced from Kaggle, available at https://www.kaggle.com/datasets/senapatirajesh/netflix-tvshows-and-movies?select=NetFlix.csv. This dataset offers a comprehensive view of TV shows and movies on Netflix up to the beginning of 2021, comprising 7787 records and 12 variables. The dataset includes information such as show type, title, director, cast, country of origin, release date, rating, duration, genre, and a brief description. I selected this dataset due to its relevance to current entertainment trends and the potential to uncover patterns in Netflix content consumption over the years.

2. **Dataset Structure and Preprocessing:**

The dataset is well-structured with 12 columns, each providing crucial information about Netflix content. To prepare the data for exploration, I addressed mixed date formats, filtered entries in the date_added column, replaced durations falling between 1-10 minutes, and substituted missing values in the rating column with 'Unknown.' Additionally, a new column was created to extract the airing year from the date_added column. These steps ensured a clean and consistent dataset for exploration.

3. **Key Visualization:**

One compelling figure from my exploratory analysis is a bar plot illustrating the distribution of Movies and TV Shows on Netflix from 2005 to 2021. The plot highlights a consistent increase in the count of TV Shows over the years, contrasting with a decline in the count of movies after 2017. This visual emphasizes the changing landscape of Netflix content, shedding light on the platform's shifting focus or audience preferences. This figure serves as a starting point for deeper investigations into the reasons behind these trends and their implications for content creators and viewers.



This exploration provides valuable insights into the dynamics of Netflix content, showcasing the platform's evolution over the years. The visualization sparks further questions about the factors influencing these trends, encouraging a more in-depth analysis of viewer preferences and industry shifts.

Reply

Attach

Cancel Post Reply

Dev Bhartra (https://northeastern.instructure.com/courses/170748/users/200511)

Feb 8, 2024

1. The dataset: The data I have used is scraped from https://h1bdata.info. This website indexes th.

1. The dataset:

The data I have used is scraped from https://h1bdata.info. This website indexes the Labor Condition Application (LCA) disclosure data from the United States Department of Labor (DOL). I extracted the data from the website with a simple python script. For brevity, I have filtered my data to only contain information pertaining to the salaries of individuals on the H1B work visa in the United States, who are working out of the city of San Francisco, California.

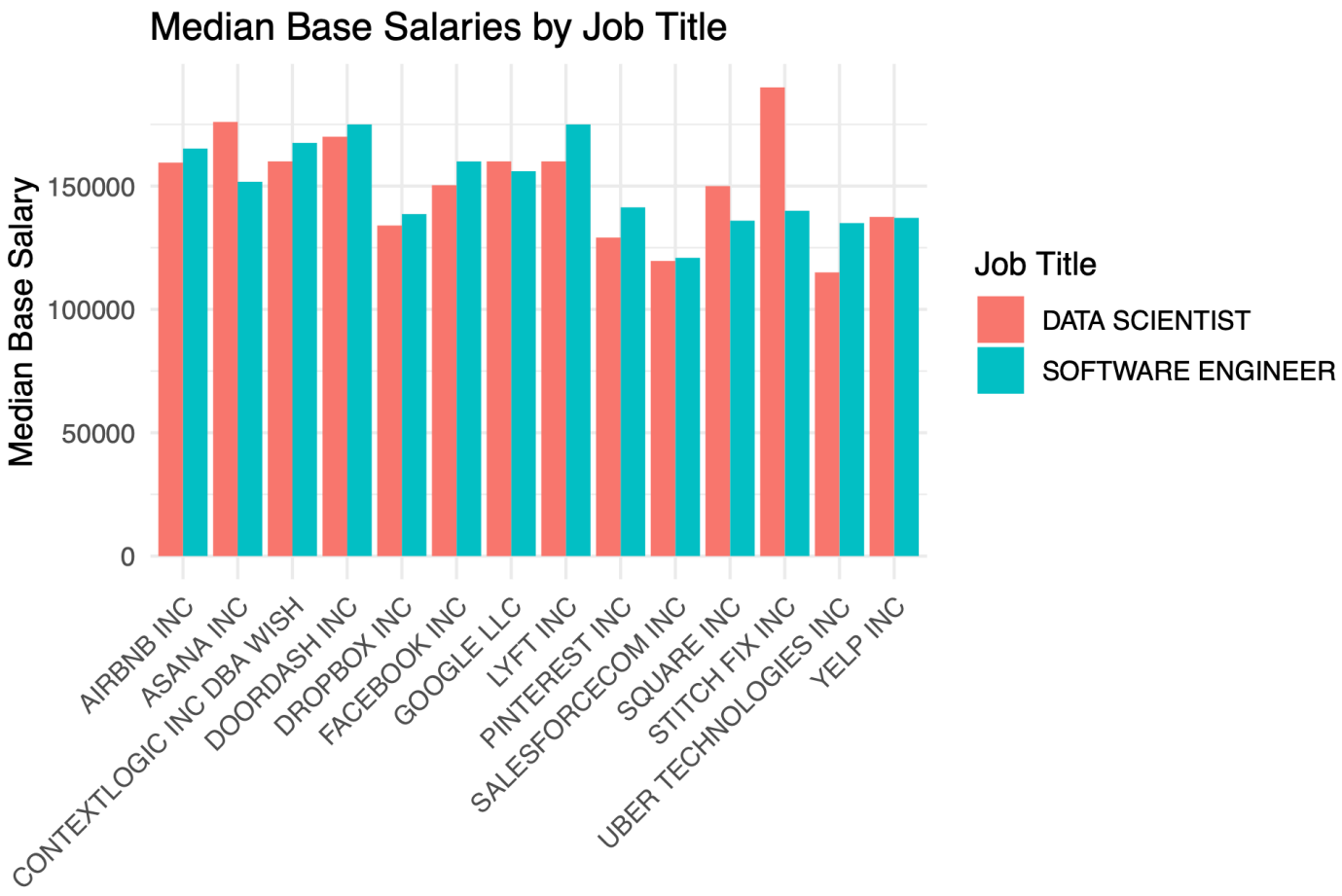
The two relevant job roles I have considered here are Software Engineer and Data Scientist.

The reason I chose this data is because it is highly relevant not only to my own career choices after I graduate from this Masters degree, but also for multiple other students in a similar position as me. By getting an understanding of the underlying stats of working in the united states, I wanted to be able to make informed decisions in future career choices. The information that can be extracted from an analysis like this is not easily available online, with popular websites like Levels.fyi not allowing users to view specifically H1B data.

2. Structure:

The raw 'untidy' data contains data points which I found important such as such as Employer, Job Title, Base Salary, Location, Work Visa Submission Date, Start Date and more. There are in total 4.8 million records at the point of writing this post, with new records added each day. I preprocessed by cleaning some missing data points. I additionally performed basic date format transformations and for visualizing data trends, extracted the year from dates as well.

3. 1 Interesting figure:



Here we see the very comparable base salaries offered to H1B candidates at top employers when comparing the roles of data scientists vs Software Engineers. These top employers pay their employees between 120 - 150k USD. Some companies have the data scientists earning slight more than software engineers, while others have the software engineers earning slightly more. Such information can be used to make smart job applications to specific companies to maximize the compensation packages that job seekers can earn, and also provide a factual backing to salary negotiations.

Edited by **Dev Bhartra (https://northeastern.instructure.com/courses/170748/users/200511)** on Feb 8 at 3:15pm

Reply

Attach

Cancel Post Reply

Raghul S (https://northeastern.instructure.com/courses/170748/users/201757)

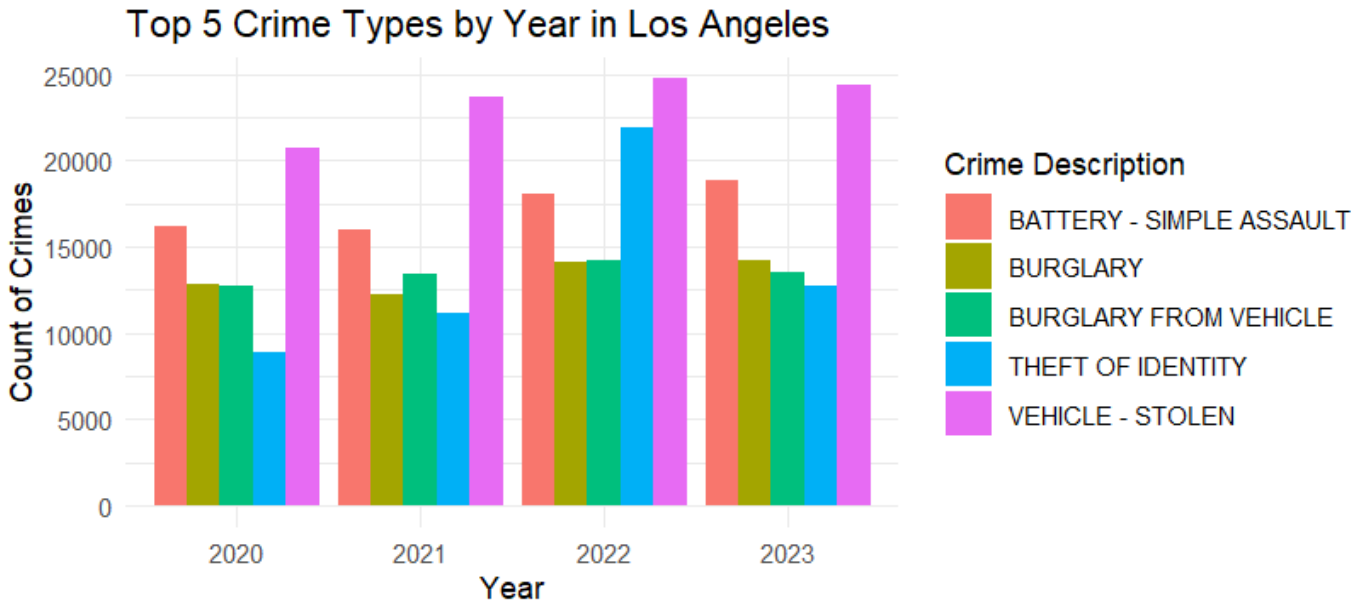
Feb 8, 2024

The dataset I chose to explore is the comprehensive crime data provided by the Los Angeles Polic.

The dataset I chose to explore is the comprehensive crime data provided by the Los Angeles Police Department, detailing reported incidents from 2020 to present, accessible via the Los Angeles Open Data Portal (**Crime Data from 2020 to Present** (https://catalog.data.gov/dataset/crime-data-from-2020-to-present)). This dataset was chosen for its extensive coverage of various crime types and the insights it can potentially offer into public safety and urban dynamics. My visualization aimed to explore trends in crime over the years, with a particular interest in understanding which types of crimes were most prevalent and if there were any notable changes in their frequency.

The dataset from the Los Angeles Police Department charts detailed crime incidents, encompassing the nature of the offense, data and time of the crime, the victim and their demographics, types of crimes, and the precise locations of each event. To prepare this data for examination, I refined it by rectifying inaccuracies, particularly in the geographical information, which involved removing records with implausible coordinates. Uniformity was brought to textual data, capitalizing all letters for consistency, and dates and times were converted into a standardized format to facilitate easier analysis. Categorization of crime types was refined, and redundant columns were removed to streamline the dataset for visualization. These pivotal preprocessing steps were carefully executed to ensure the dataset's reliability, paving the way for a robust analysis of Los Angeles' crime trends.

The below figure, a bar chart depicting the top five crime types by year, reveals intriguing patterns. For instance, 'Theft of Identity' and ‘Vehicle - Stolen’ shows a marked increase, while 'Battery - Simple Assault' and 'Burglary from Vehicle' hold steady prominence. This visualization highlights the evolving landscape of crime in Los Angeles, providing a platform for further discussion on crime prevention and resource allocation. This observation underscores the evolving nature of crime. Such a trend calls for an adaptive approach in public safety strategies, highlighting the necessity of continuous monitoring and proactive measures in law enforcement.



Reply

Attach

Cancel Post Reply

Kalli A. Hale (*She/They*) (https://northeastern.instructure.com/courses/170748/users/251230)

Feb 9, 2024

Haunted Places in the United States Background: Since December 1994, TheShadowLands.net h

Haunted Places in the United States

Background:

Since December 1994, TheShadowLands.net has created space for speculatory information on the

supernatural. In particular, its **Haunted Places Index (https://www.theshadowlands.net/places/)** has allowed users to self-report local hauntings, with tens of thousands of unique submissions from people and places all around the world.

Compiled by Timothy Renner, and shared by the R for Data Science Tidy Tuesday project, **this data set explores nearly 11,000 reportedly haunted places in the United States. (https://github.com/rfordatascience/tidytuesday/tree/master/data/2023/2023-10-10#haunted-places-in-the-united-states)** In addition to a description of the alleged paranormal activity, it includes such variables as city, state, named location, and coordinate points for both location and city.

Manipulation and Analysis:

To make the latitude and longitude columns consistent with their counterparts in the "maps" package for different forms of spatial visualization, I renamed the latitude and longitude variables to "lat" and "long", respectively. Additionally, two redundant variables were removed: "country", which is unnecessary because we can assume here that all observations occur in the United States, and "state_abbrev", as the state itself is already given in the "state" column. Additionally, I create a new category called "region" which categorizes the state in which each haunting takes place into one of eight culturally significant regions.

The regions carved out for the problem were chosen carefully for their cultural significance. For instance, the "Deep South" refers to states in which American chattel slavery was deeply entrenched, and therefore potent sites of mass death, torture, and terrorism of enslaved African peoples, a cultural specter of quietly accepted horror which still looms in the collective imagination through Southern gothic literatures and aesthetics. The entire United States is a colonial project founded on the genocide of Indigenous peoples in order to facilitate land theft; a fact that is underscored and preserved within the colonial imagination with many infamous descriptions of the ghosts of Indigenous people frequently evidenced in this data set.

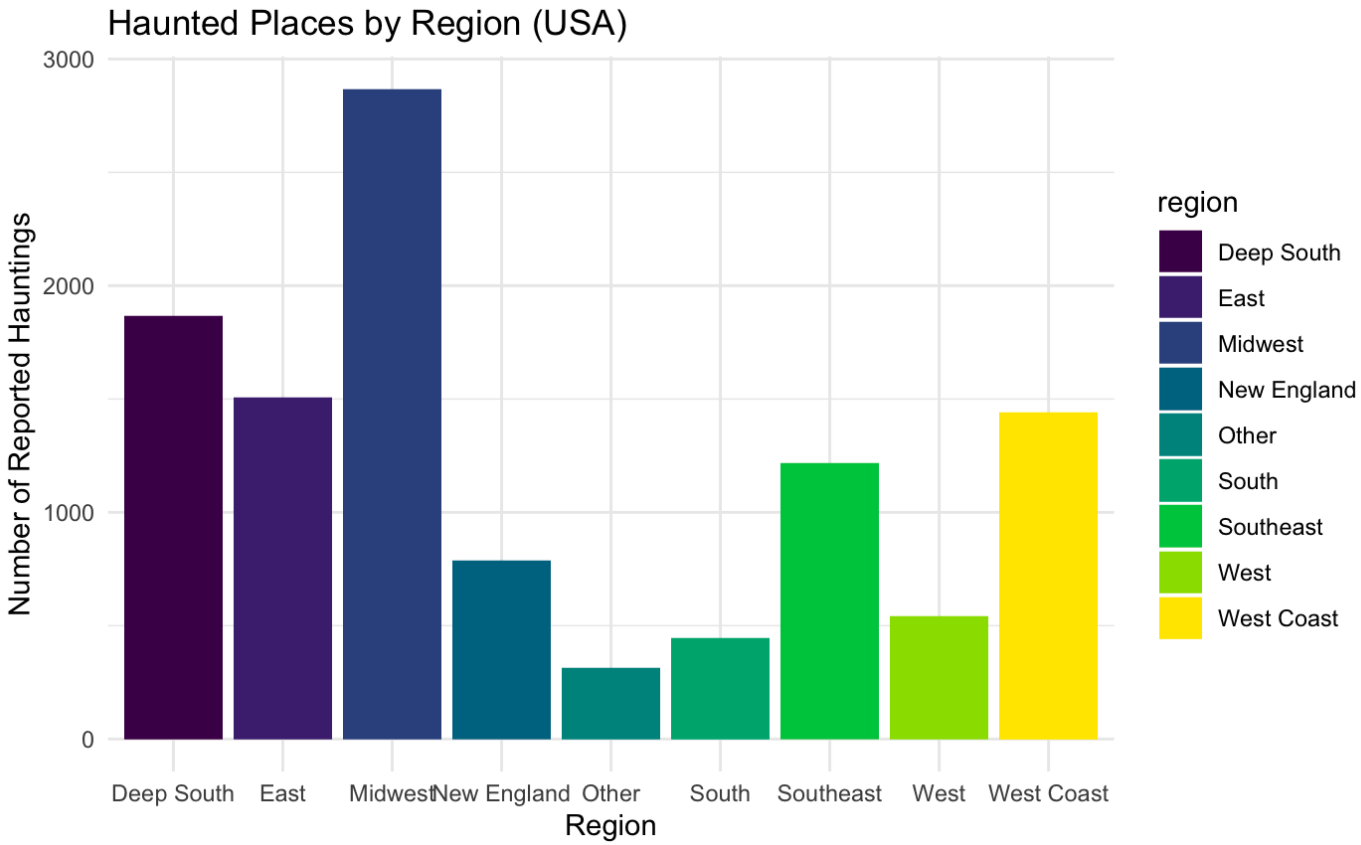
These subgroups also attempt to reflect thoughts about areas with notably disparate population densities. For example, "West," a category which here holds double the number of states as the contiguous "West Coast" category, has far fewer responses than the latter, which contains only California, Oregon, and Washington.

The Midwest is the largest category with the most states, so it isn't surprising that we see so many reported hauntings there. Interestingly, the "Deep South" category is fairly high in terms of reported hauntings! However, that does include Texas, one of the largest and most populous states in the country.

Final Commentary:

For a more in-depth analysis, it would be interesting to search for keywords in the haunting

description to find mentions of specific ghostly archetypes (i.e., "girl" or a collection of words like c("lover", "boyfriend", "husband", "wife", "girlfriend"), etc., and visualize their frequencies, or to look at places of historical significance and visualize purported hauntings. Breaking this data set down more thoroughly with respect for differences in overall population and concentrations of population density in an attempt to identify disproportionately "haunted" places in the United States and analyzing the cultural/historical significance of these places would be an interesting anthropological exercise. Also--are there more ghosts in cities, or on the countryside? Would it make sense for cities, given that there is more foot traffic? Or might we see that, in the aggregate, the relatively more isolated nature of rural landscapes and country living inspire more imagined spiritual or ghostly encounters?



Edited by **Kalli A. Hale (https://northeastern.instructure.com/courses/170748/users/251230)** on Feb 13 at 11:37am

Reply

Attach

Cancel Post Reply

Aditya Vyas Gurnani (https://northeastern.instructure.com/courses/170748/users/206276)

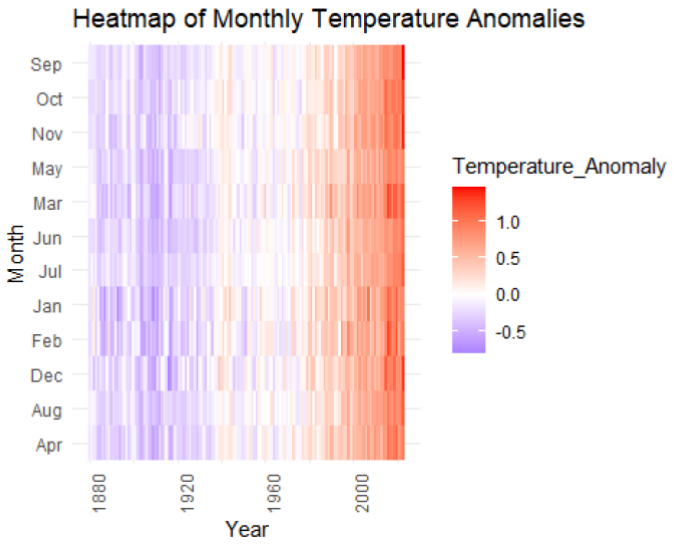
Feb 11, 2024

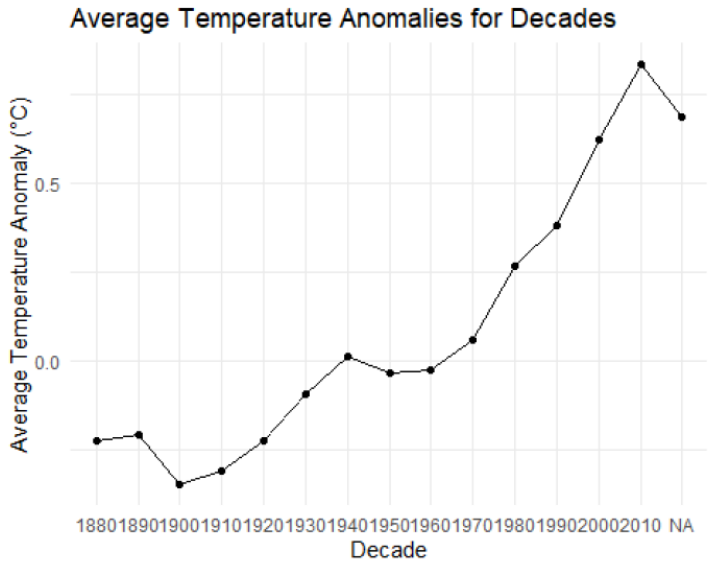
The "Global Surface Temperature Change" dataset was provided by NASA's Goddard Institute for .

The "Global Surface Temperature Change" dataset was provided by NASA's Goddard Institute for Space Studies (GISS). This dataset, a comprehensive record of global surface temperature anomalies measured in degrees Celsius against a mid-20th-century baseline, was chosen for its relevance to understanding climate change trends over time. My interest in this dataset was driven by a desire to explore the progression of global warming and identify any patterns or anomalies in temperature changes across different seasons and decades.

The dataset consists of annual and monthly temperature anomalies from the late 19th century to the present. Variables include the observation year, monthly anomalies from January to December, and seasonal anomalies categorized as Winter (DJF), Spring (MAM), Summer (JJA), and Autumn (SON), alongside the annual mean anomaly (J-D). Preprocessing steps involved replacing missing values denoted by "***" with NA, ensuring all numerical columns were correctly formatted for analysis, and renaming columns for clarity.

The heatmap of monthly temperature anomalies and the line graph of decadal average temperature anomalies, two visualizations made with this dataset, offer startling new insights. The heatmap, which uses warmer colors to indicate higher temperature anomalies, effectively depicts the trend toward warmer temperatures over the past few years. The decadal line graph quantitatively supports the consensus on global warming by demonstrating a definite rising pattern in global temperatures.





These explorations not only emphasize the critical nature of long-term climate trends but also highlight the importance of datasets like GISS's in informing policy and public awareness around climate change. The analysis underscores the urgent need for action to mitigate global warming effects, reflecting a broader scientific understanding of climate change dynamics.

Reply

Attach

Cancel Post Reply

Pranit Brahmbhatt (*He/Him*) (https://northeastern.instructure.com/courses/170748/users/200535)

Feb 11, 2024

Exploring Tech Layoffs Dataset 1. Dataset Description: I chose to analyze the Tech Layoffs dataset

Exploring Tech Layoffs Dataset

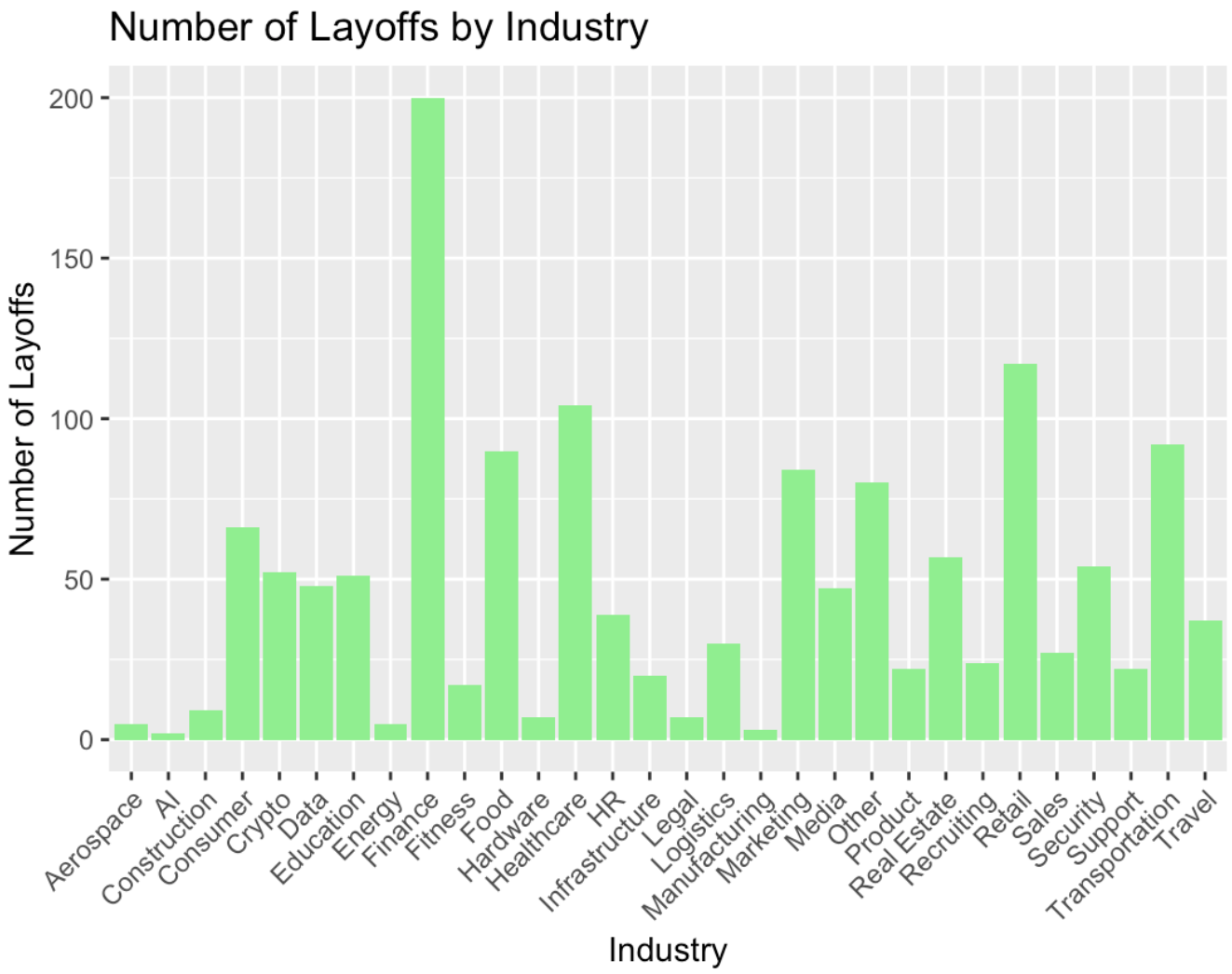
1. Dataset Description: I chose to analyze the Tech Layoffs dataset sourced from Kaggle (https://www.kaggle.com/datasets/ulrikehe rold/tech-layoffs-2020-2024/data (https://www.kaggle.com/datasets/ulrikeherold/tech-layoffs-2020-2024/data)) for my assignment. This dataset offers insights into tech layoffs spanning the years 2020 to 2024. I selected this dataset due to its relevance in understanding the ongoing dynamics of layoffs within the technology industry. My aim was to explore various aspects such as the distribution of layoffs across industries, company stages, and geographical locations, and see how critically the market is getting hit to the extent that the startups and companies are closing more rapidly than ever before by laying off their entire workforce.

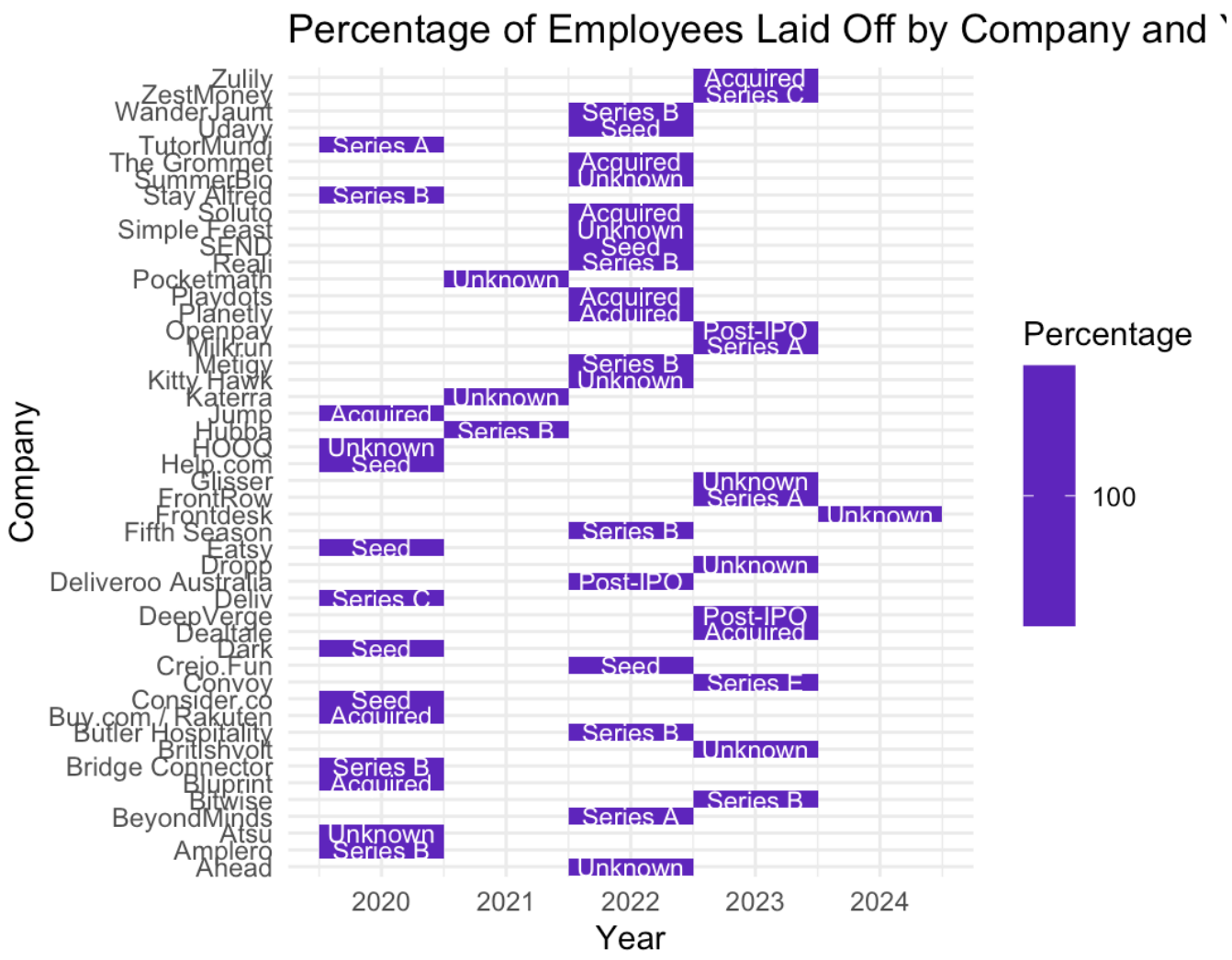
2. Dataset Structure and Preprocessing: The dataset was already in a tidy format, meeting the criteria where each column represents a variable, each row represents an observation, and each cell contains a single value. It includes variables such as Company, Location_HQ (Location of their Headquarter), Country, Continent, Laid_Off (Number of employees laid off), Date_Layoffs (The date a recent layoff for a company), Percentage (Percentage of employees laid off from a company), Company_Size_before_layoffs, Company_Size_after_layoffs, Industry, Stage (Recent Funding Stage for a company), Money_Raised_in_$_mill, Year, lat (latitude co-ordinates for a company), long (longitude co-ordinates for a company). No preprocessing was required as the dataset was clean and ready for analysis.

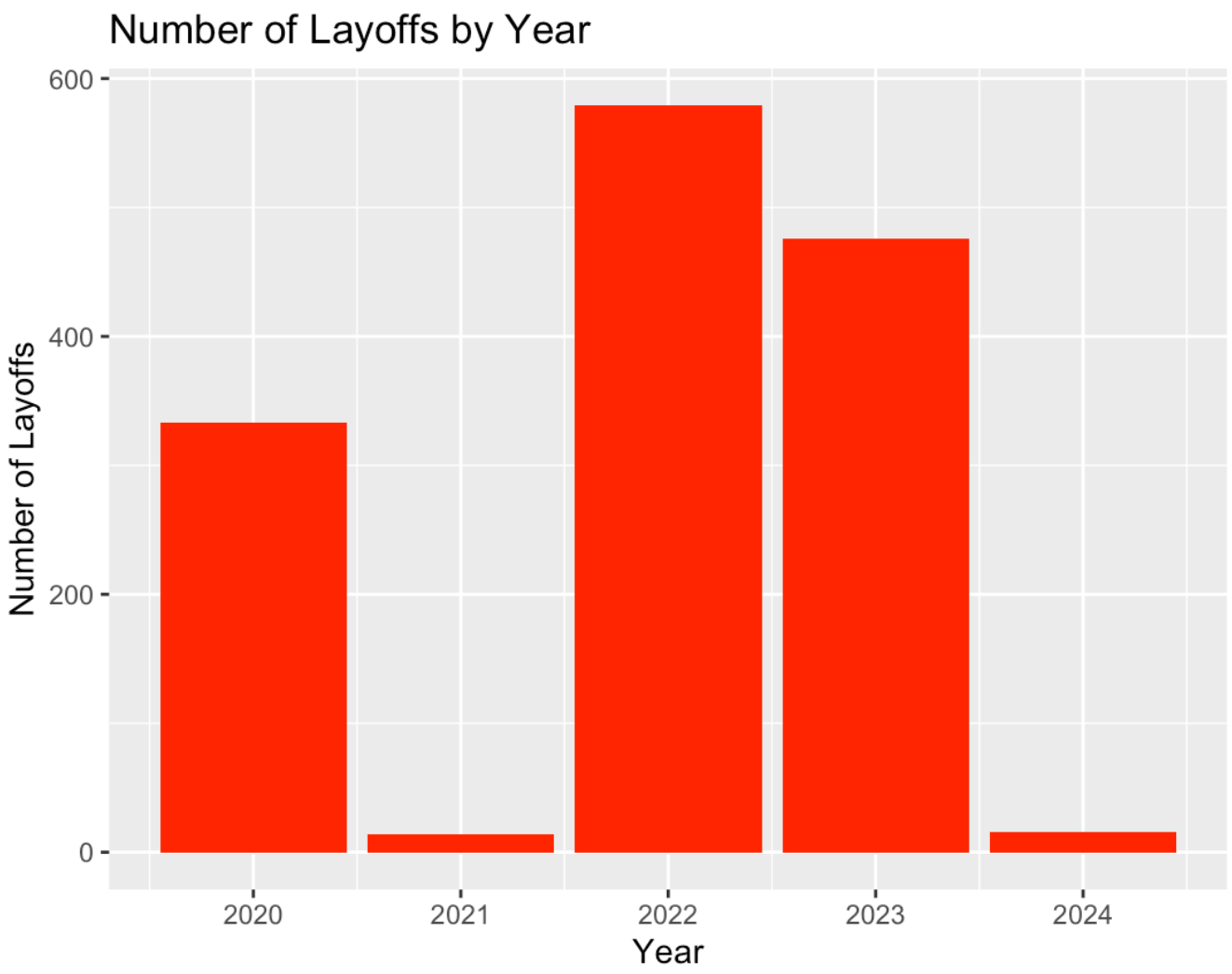
3. Visualization and Analysis: I conducted exploratory analysis and created several visualizations to understand the trends within the dataset:

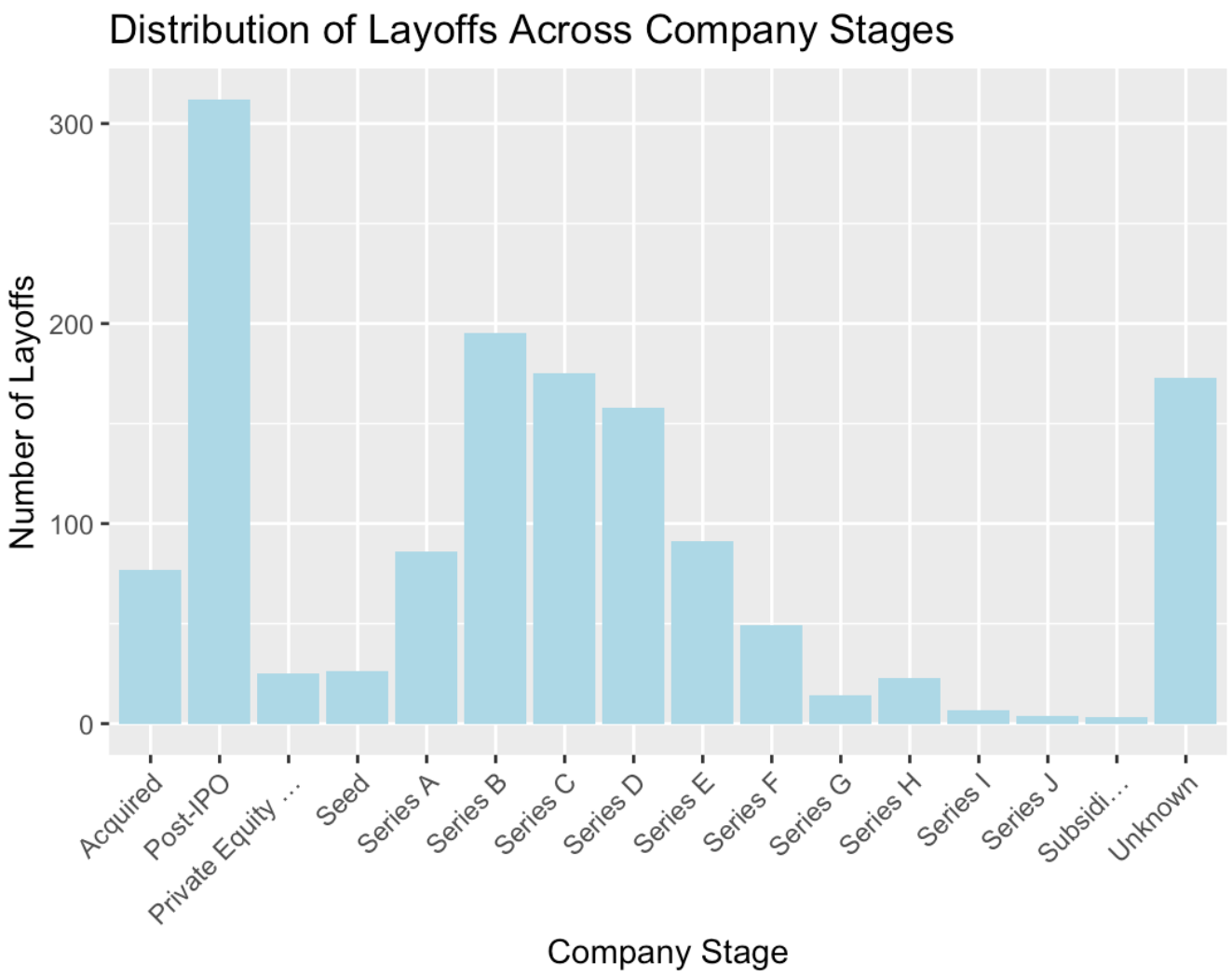
- **Number of Layoffs by Industry:** This visualization allowed me to identify which industries experienced the highest number of layoffs.
- **Distribution of Layoff Percentages:** Examining the distribution of layoff percentages provided insights into the severity of layoffs across companies.
- **Distribution of Layoffs Across Company Stages:** Analyzing layoffs based on company stages helped in understanding how layoffs vary based on the funding stage of companies.
- **Number of Layoffs by Year:** This visualization depicted the trend of layoffs from 2020 to 2024.
- **100% employees laid of by a company every year:** An analysis of companies laying off 100%

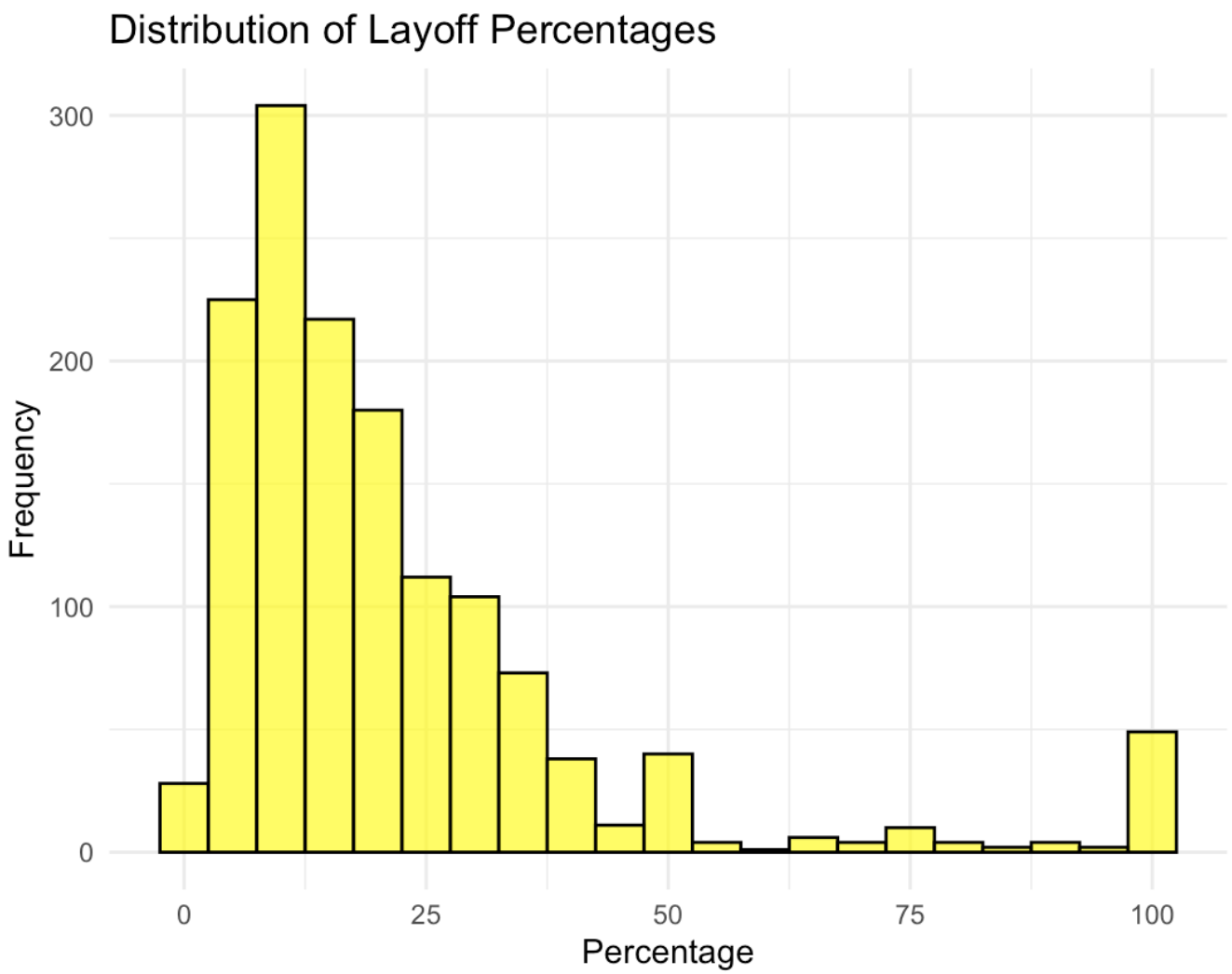
of its workforce, meaning the company is now closed and no longer functioning entirely. Moreover, to gain an in-depth analysis, I have added the Stage variable to help us know at what stage the company had to be shut down. One noteworthy observation was the correlation between company stage and layoffs. Interestingly, many companies that faced closure were at early funding stages, suggesting potential vulnerabilities in the startup ecosystem and how a market disruption can severely impact the startup bubble.




Percentage of Employees Laid Off by Company and `
Company
Year
2020
2021
2022
2023
2024
Percentage
100
Zulily
ZestMoney
WanderJaunt
Udavy
TutorMundi
The Grommet
SummerBio
Stay Alfred
Soluto
Simple Feast
SEND
Reali
Pocketmath
Playdots
Planetly
Openpay
Milkrun
Metigy
Kitty Hawk
Katerra
Jump
Hubba
HOOQ
Help.com
Glisser
FrontRow
Frontdesk
Fifth Season
Eatsy
Dropp
Deliveroo Australia
Deliv
DeepVerge
Dealtale
Dark
Crejo.Fun
Convoy
Consider.co
Buy.com / Rakuten
Butler Hospitality
Britishvolt
Bridge Connector
Bluprint
Bitwise
BeyondMinds
Atsu
Amplero
Ahead
Acquired
Series C
Series B
Seed
Series A
Acquired
Unknown
Series B
Unknown
Acquired
Seed
Series B
Unknown
Acquired
Acquired
Post-IPO
Series A
Series B
Unknown
Unknown
Acquired
Series B
Unknown
Seed
Unknown
Series A
Unknown
Series B
Seed
Unknown
Post-IPO
Series C
Post-IPO
Acquired
Seed
Seed
Series F
Acquired
Series B
Unknown
Series B
Acquired
Series B
Series A
Unknown
Series B
Unknown


Number of Layoffs by Year
600
400
200
0
Number of Layoffs
2020
2021
2022
2023
2024
Year


Distribution of Layoffs Across Company Stages
Number of Layoffs
300
200
100
0
Acquired
Post-IPO
Private Equity …
Seed
Series A
Series B
Series C
Series D
Series E
Series F
Series G
Series H
Series I
Series J
Subsidi…
Unknown
Company Stage



Reply

Attach

Cancel Post Reply

-

Teja Ramana Modukuru (https://northeastern.instructure.com/courses/170748/users/217600)

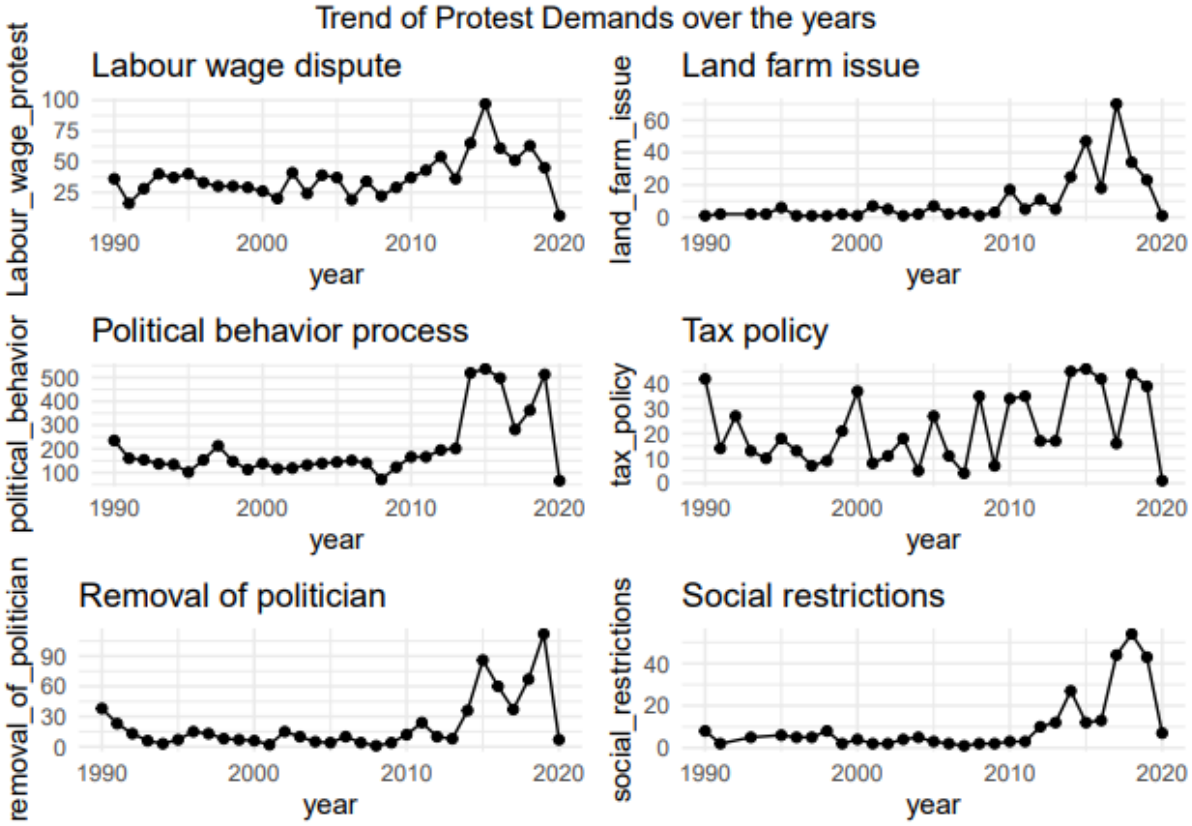
Feb 13, 2024

The Mass Mobilization Project dataset is a comprehensive collection of data on citizen movements

The Mass Mobilization Project dataset is a comprehensive collection of data on citizen movements against governments. It covers 162 countries from 1990 to March 2020. The data includes both instances of anti- and pro-regime protests at the city level with daily resolution. It also records up to seven types of government responses, including accommodation, arrests, beatings, crowd dispersal, ignore, killings, and shootings.
The data is coded based on news reports obtained from AP, the AFP, and BBC Monitoring. The project is sponsored by the Political Instability Task Force (PITF). You can find the Mass Mobilization dataset at the official project website and also at the Harvard Dataverse. (**https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/HTTWYL (https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/HTTWYL)**)
This dataset was chosen because it offers a comprehensive view of citizen movements against governments across a wide range of countries and over a significant period of time. The detailed categorization of protests and government responses in the dataset provides a unique opportunity to explore various aspects of these movements.
The specific questions I wanted to explore in my visualization are:
1. Trends of Protest Demands over the Years
2. Trends of Number of Protests over all the Countries
3. Trends of Durations of Protests in all the Continental Regions
4. Type of state responses over protestors of various countries
5. Variation in the State's Response over the various demands of protestors

The Mass Mobilization Project dataset is structured as a table with each row representing a unique protest event. Here are the variables in the dataset:
id: A unique identifier for each protest event.
country: The country where the protest took place.
ccode: The Correlates of War numeric code for the country.
year: The year of the protest.
region: The region of the world where the protest took place.
protest: A binary variable indicating whether a protest took place.
protestnumber: A count of protest events for the country in that year.
startday, startmonth, startyear: The date when the protest started.
endday, endmonth, endyear: The date when the protest ended.
protesterviolence: A binary variable indicating whether the protesters were violent.
location: The city or region of the protest.
participants_category: A categorical variable indicating the size of the protest.
participants: The estimated number of participants.
protesteridentity: The identity of the protesters.
protesterdemand1, protesterdemand2, protesterdemand3, protesterdemand4: The demands of the protesters.

stateresponse1, stateresponse2, stateresponse3, stateresponse4, stateresponse5, stateresponse6, stateresponse7: The responses of the state to the protest.
sources: The sources of the information.
notes: Additional notes about the protest event.
This structure allows for a detailed analysis of protest events, including when and where they occurred, who was involved, what the protesters demanded, and how the state responded.
The preprocessing needs for the Mass Mobilization Project dataset include cleaning missing or inconsistent data, transforming data into a suitable format for analysis, and tidying the dataset to ensure a structured format, In the preprocessing stage, the columns with more than 50% missing values were removed, any remaining rows with missing values were dropped, and all commas were removed from the 'protesterdemand1' column to prevent any potential issues during analysis. These steps helped ensure the dataset was clean and ready for further analysis.



Insights & Conclusion
While the first viewing of the data hints at a sudden peak from the year 2013, it is most likely due to better form of data collection techniques by the authorities over the years, namely - digitization, collaboration between multiple departments, and so on.
Tax Policy - The only chart that stands out as there have been protests registered throughout the decades. The seasonality in sudden peaks and troughs can be noticed as those years where the majority of elections or crucial budgets must have been presented over the various regions. However, post-2012, there have been years of persistent protests regarding Tax Policy. 2017 being a lone case,

all other years have seen a consistent amount of protests regarding Tax Policies adopted by the government.

Social Restrictions - Protests against restrictions laid down by society or governments on various social practices have been demonstrated for centuries. However, with the advent of social media, we can see that the number of protests on social restrictions has considerably increased as more and more people are aware of the various topics of discussion.

These insights highlight the power of social movements and the impact of government policies and societal norms on public sentiment

In conclusion, different governments throughout the world have different responses to demonstrations, with ignoring being the most typical one. Across all continents, the majority of protests last 0 to 30 days, with few lasting longer. Additionally, the demands of the protests differ in terms of content, with tax policy and the impeachment of officials serving as two recurring themes. The data as a whole indicates that demonstrations over diverse demands have been a persistent trend, and that state officials have responded to them in a combination of tolerant and aggressive ways.

Edited by Teja Ramana Modukuru (https://northeastern.instructure.com/courses/170748/users/217600) on Feb 15 at 9:43pm

Reply

Attach

Cancel Post Reply

Mohan Vilasrao Bhosale (https://northeastern.instructure.com/courses/170748/users/295364)

Feb 13, 2024

Flash Paper Describe the dataset and where it comes from (making sure to cite the data source).

Flash Paper

1. **Describe the dataset and where it comes from (making sure to cite the data source). Explain why you chose this dataset and what questions you wanted to explore in your visualization.**

Ans:

I've selected the layoff dataset available on Kaggle. (Source: https://www.kaggle.com/datasets/swaptr/layoffs-2022) Given the current economic recession, companies are encountering financial challenges, prompting them to downsize their workforce in order to mitigate costs and maintain financial stability. Notably, Meta recently terminated 13% of its employees, totaling over 11,000 employees. I opted for this dataset to investigate recent layoffs within the tech industry, as it provides the most up-to-date information and directly impacts individuals like myself, who are pursuing graduate studies. The specific questions I aimed to address through visualization were:

1. Which industry is experiencing the greatest impact from layoffs?
2. What is the breakdown of laid-off employees by region?
3. How has the trend of layoffs evolved in previous years?
4. Are layoffs still occurring in 2024, and if so, which sectors are predominantly affected?

The dataset is recently updated on Kaggle and has the data from ***March 2020*** to ***Jan 2024***.

---

2. **Describe the structure of the dataset and the variables of interest. Describe any preprocessing needs (tidying, cleaning, transformation, etc.) and describe the steps you took to perform the preprocessing.**

Ans:

The layoff dataset comprises 3,319 rows and 9 variables. Below are descriptions of some key variables:

- ***company*** : Name of the company
- ***location*** : Location of company headquarters
- ***industry*** : Industry of the company
- ***total_laid_off*** : Number of employees laid off
- ***percentage_laid_off*** : Percentage of employees laid off
- ***date*** : Date of layoff

- ***stage*** : Stage of funding
- ***country*** : Country
- ***funds_raised*** : Funds raised by the company (in Millions $)

To enhance the dataset's tidiness and utility for analysis, the following steps were undertaken:

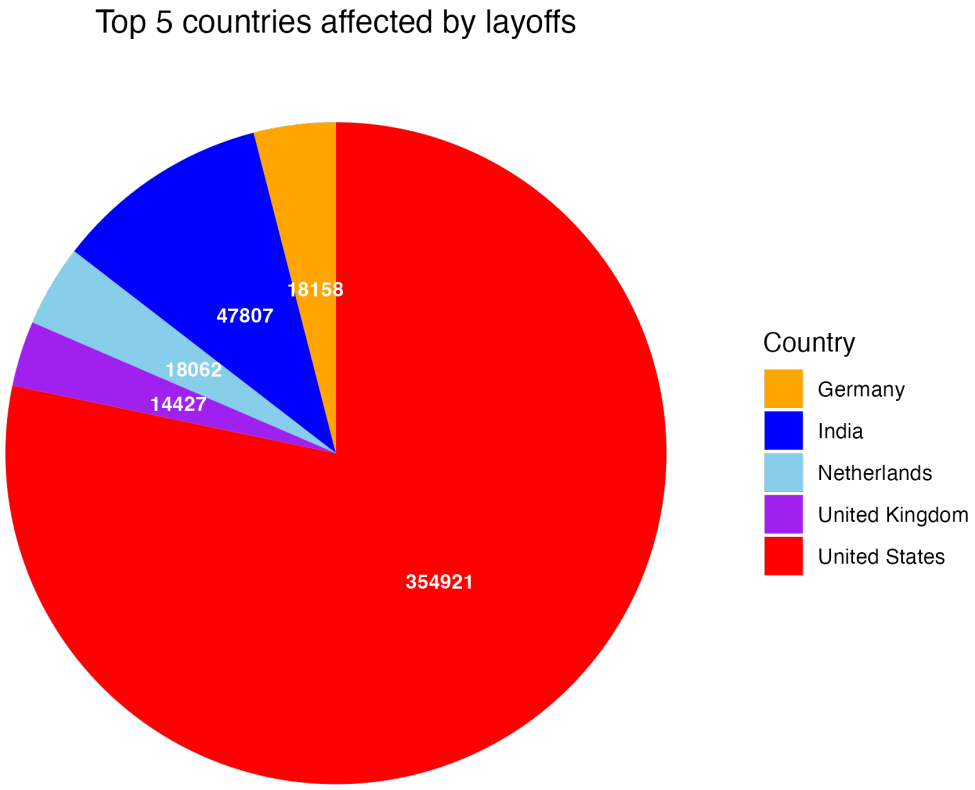
1. Initial assessment of the dataset's dimensions, including the summary of dataset.
2. Identification and reporting of missing values (NAs) across all columns, such as company, date, stage, etc.
3. Imputation of missing values in the 'total_laid_off' column by assigning a value of 1, indicating a single employee laid off, to retain the row's relevance.
4. Assigning an "unknown" value to missing values in the 'industry' and 'location' columns.
5. Conversion of the data type of 'total_laid_off' to integer format.
6. Conversion of the data type of the 'date' variable to the date format for consistency and ease of analysis.

---

3. **Present at least 1 figure that is interesting to you and describe your observations and any key takeaways from the visualization and your exploration of the dataset.**

Ans:

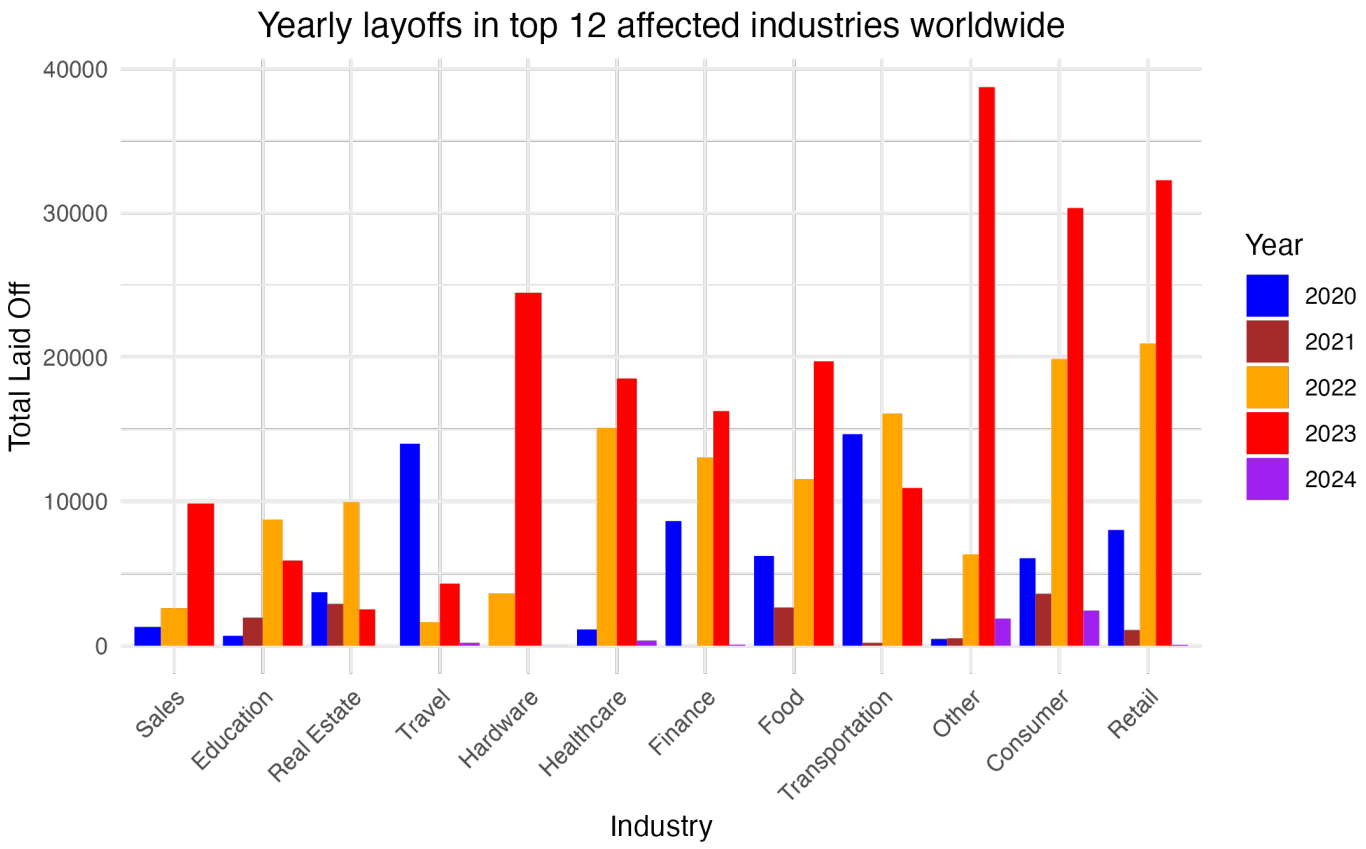
1. *Total number of employees laid off in the top 5 countries affected by the layoffs:*



Key Points:

- The pie chart above illustrates the distribution of the top 5 countries in terms of laying off employees up to January 2024.
- The United States emerges as the most affected country globally, with 354,921 employees laid off.
- Following the United States, India ranks second with a total of 47,807 employees affected.
- This trend is further followed by Germany, the Netherlands, and the United Kingdom, respectively.

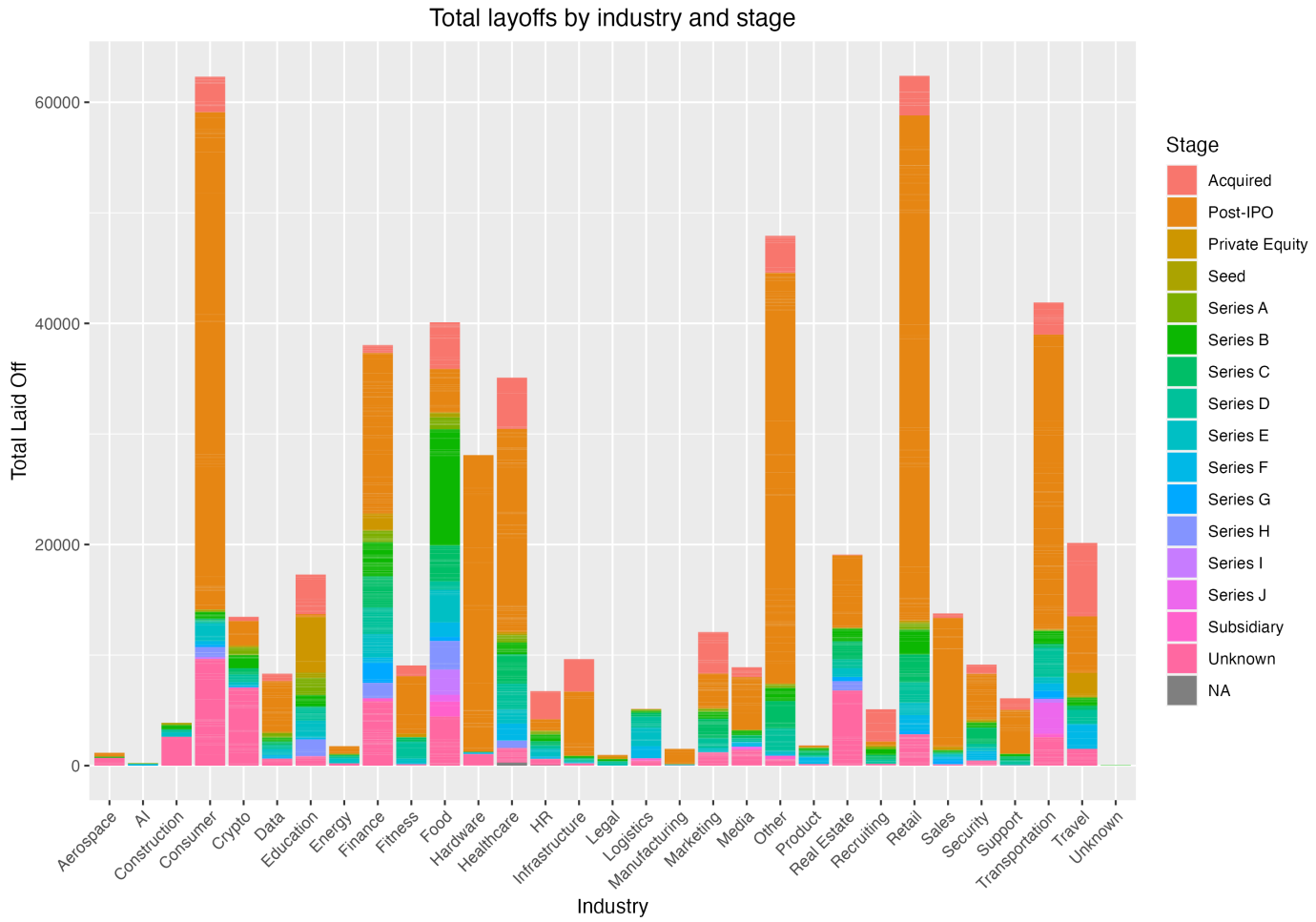
2. Year-on-year layoffs in the industries affected by the layoffs worldwide:



Key Points:

- The data indicates that 2023 marked the peak year for layoffs across various industries.
- Both the transportation and education sectors demonstrated significant recovery, experiencing fewer layoffs in 2023 compared to 2022.
- Conversely, industries such as consumer goods, healthcare, and travel witnessed notable layoffs in the initial month of 2024.
- Notably, the hardware industry stands out as one of the few sectors that did not execute any layoffs in 2020.

3. *Total layoffs by industry and stage of the company*



Key Insights:

- Post-IPO and Series B companies are responsible for the majority of large-scale layoffs, indicating that several large-scale companies have conducted significant workforce reductions.
- The Consumer and Retail sectors maintain their prominence in the layoff trend through January 2024.
- Layoff rates peaked in 2020, declined in 2021, and surged in 2022.
- The trend of layoffs persists into January 2024.
- The **AI industry** demonstrates comparatively fewer layoffs. However, this does not necessarily imply it is a secure career path; rather, the industry is new and evolving, which is the reason for fewer layoffs.
- Product, Manufacturing, Energy, and Aerospace industries exhibit the lowest levels of impact among the mass layoffs.

Thank you!

- Mohan Bhosale

Reply

Attach

Cancel Post Reply

-

Divya Chenduran (*He/Him*) (https://northeastern.instructure.com/courses/170748/users/261307)

Feb 13, 2024

1. The dataset I looked at was obtained from DATA.GOV and contains information on every cancer

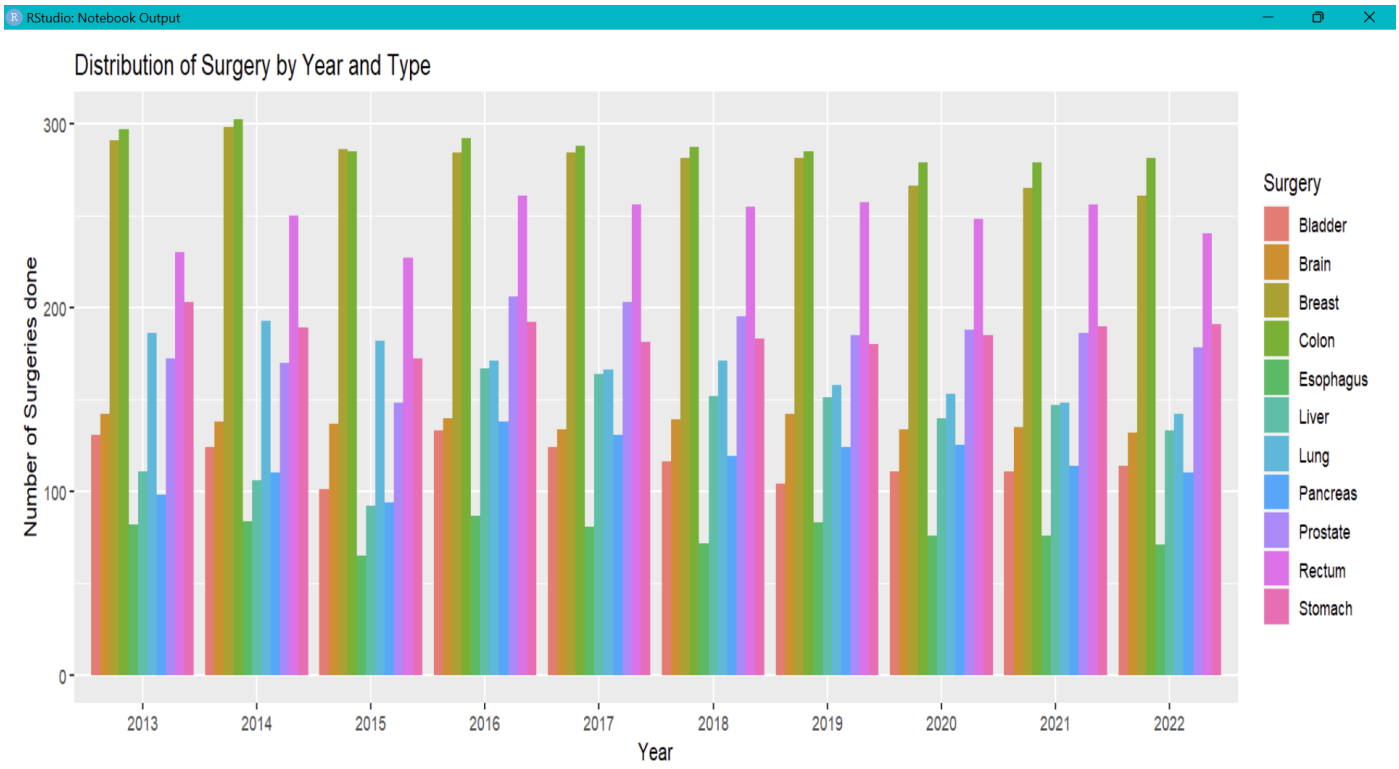
1. The dataset I looked at was obtained from DATA.GOV and contains information on every cancer operation that has ever been done in the state of California. To access the dataset, click on this link: https://catalog.data.gov/a3f18. I was curious to find more about the various cancer surgery procedures that are done every year and the hospitals that conduct them. By examining the data, I was able to determine which counties and hospitals conducted the most surgeries of cancer procedures.

2. The information included the various surgical procedures—such as brain, colon, and prostate—as well as the annual breakdown of the number of cases for each type of cancer surgery carried out by various institutions in various counties. We must rename the columns and remove the empty values from this dataset before using it. I changed the incorrect column name to the right name and used "complete.cases" to remove the null entries.

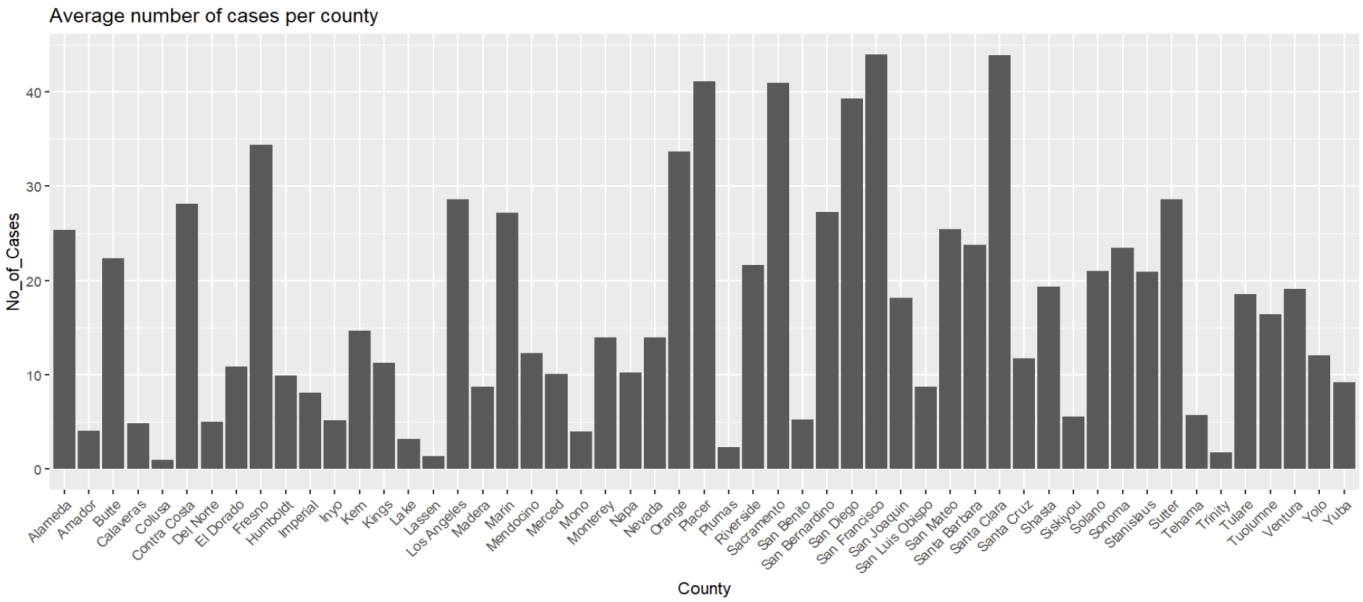
Structure of the dataset-
Year - from 2013 to 2022 ,
County - different counties in california,
Hospital - the hospitals where the cancer surgeries were performed,

Surgery - the type of cancer,
Number of cases - the number of cases for different cancer types,
Latitude,
Longitude.

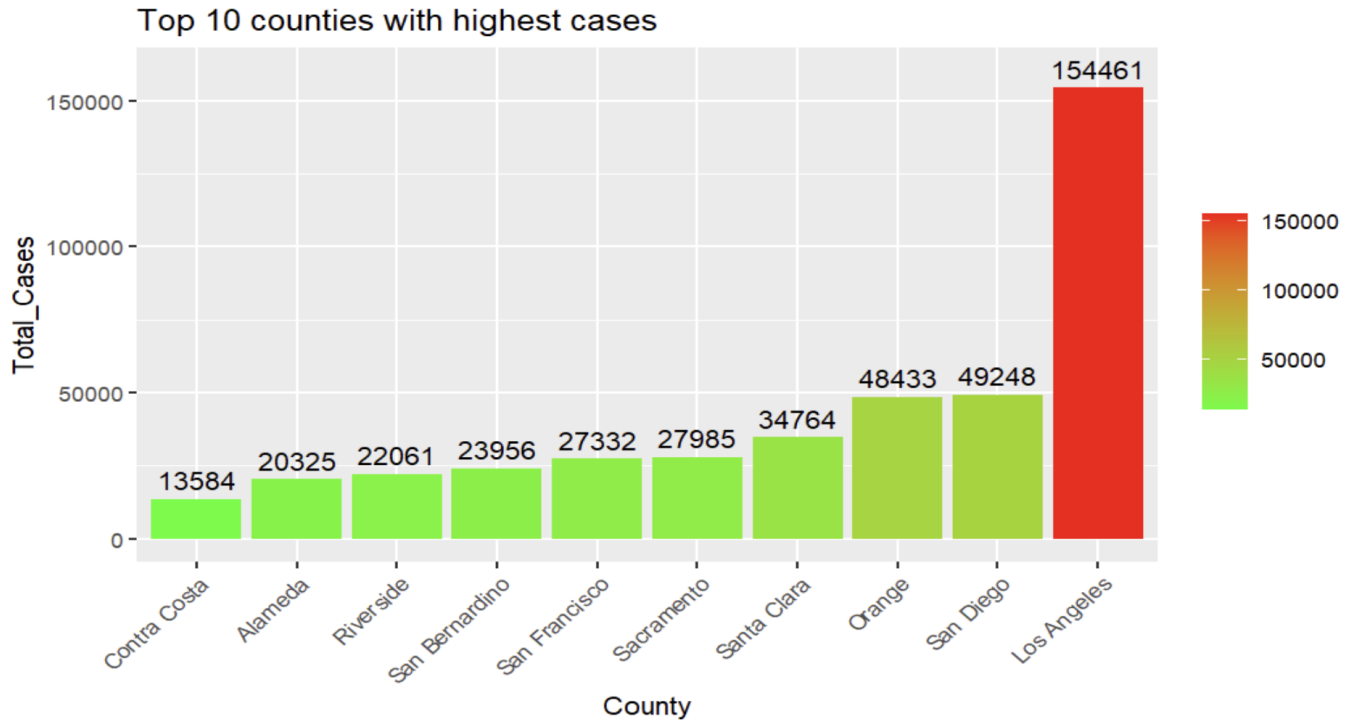
3.



This graphic shows the various cancer procedures that were done between 2013 and 2022, divided out by year. According to the depiction, breast cancer comes in second place on the graph, behind colon cancer. Additionally, we see that starting in 2016, there has been a sharp rise in the number of liver cancer cases. We may observe that the number of instances for the other operations is nearly equal.



The average number of instances for each county is displayed in the graphic above. With slightly more than 40 instances on average, "santa clara" and "san franciso" have the highest average, according to the bar graph. Colusa has the lowest average number of cases—roughly two on average.



The top ten counties with the highest number of cases are displayed in the bar graph above. San Diego comes in second with 49248 cases, with Contra Costa having 13584 cases has the fewest cases. Los Angeles is ranked #1 with 154461 cases.

Edited by **Divya Chenduran (https://northeastern.instructure.com/courses/170748/users/261307)** on

Feb 14 at 3:50pm

Reply

Attach

Cancel Post Reply

-

Yash Khare (*He/Him*) (https://northeastern.instructure.com/courses/170748/users/306131)

Feb 13, 2024

ENERGY STATISTICS DATASET INFO DATASET LINK The dataset is an excel file with 30 sheets

ENERGY STATISTICS

DATASET INFO

DATASET LINK (https://energydata.info/dataset/key-world-energy-statistics-enerdata/resource/dcda6530-8d2c-436e-9d1d-1e2809ad303e)

- The dataset is an excel file with 30 sheets all containing quality time series on supply, demand and trade for oil, gas, coal and electricity as well as information on renewable energies and CO2 emissions.
- The present Yearbook export covers 45 countries.
- Each sheet contains countries and their energy values for years ranging from 1990 to 2020 (30 years).

PROBLEM JUSTIFICATION

- **Comprehensiveness**: The dataset contains information on various energy sources including oil, gas, coal, electricity, and renewables, as well as metrics like CO2 emissions. This comprehensive coverage allows for a holistic analysis of the energy landscape.
- **Temporal Span**: With data spanning three decades (1990-2020), this dataset offers the opportunity to observe long-term trends and patterns in energy production, consumption, and

environmental impacts. Examining such trends can provide valuable insights into the evolution of the energy sector over time.

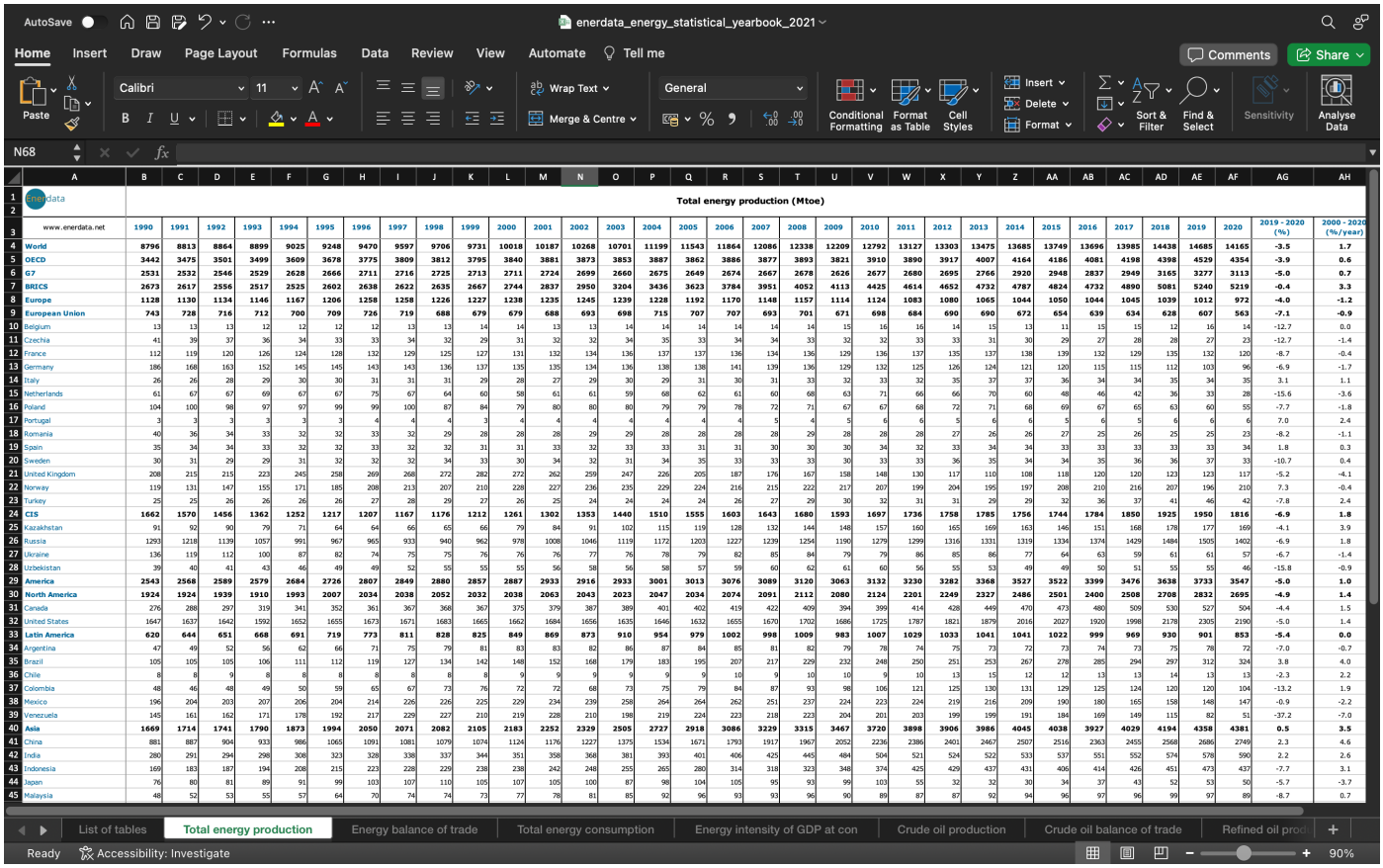
- **Geographical Coverage**: Covering 45 countries, the dataset enables comparative analysis across different regions and economies. Exploring energy-related metrics at the national level allows for understanding regional variations and disparities in energy dynamics.
- **Relevance**: Energy is a critical component of global socio-economic development, and understanding its dynamics is essential for addressing sustainability and environmental challenges. By analysing this dataset, we can gain insights into key factors influencing energy transitions, policy decisions, and environmental outcomes.

OBJECTIVES

- **Temporal Trends**: What are the long-term trends in total energy production and consumption across different countries? How have these trends evolved over the past three decades?
- **Drivers of Emissions**: What are the primary drivers behind changes in CO2 emissions? Are there notable correlations between economic growth and CO2 emissions intensity?
- **Petroleum Products Trends**: What are the long-term trends in petroleum products production and consumption? How have these trends evolved over time, and are there any notable shifts in production & consumption patterns?

DATASET STRUCTURE

- Excel contains multiple sheets with similar structure of data, i.e. country and year columns with few irrelevant rows and columns.
- Country column is under a different named column `*www.enerdata.net*` which is not located at the first row.
- Categories are present as sheet names which needs to be inserted in the data frame along with countries and years.



	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2019 - 2020 (%)	2000 - 2020 (%/year)
World	8796	8813	8864	8899	9025	9248	9470	9597	9706	9731	10018	10187	10268	10701	11199	11543	11864	12086	12338	12209	12792	13127	13303	13475	13685	13749	13696	13985	14438	14685	14165	-3.5	1.7
OECD	3442	3475	3501	3499	3609	3678	3775	3809	3812	3795	3840	3881	3873	3853	3887	3862	3886	3877	3893	3821	3910	3890	3917	4007	4164	4186	4081	4198	4398	4529	4354	-3.9	0.6
G7	2531	2532	2546	2529	2628	2666	2711	2716	2725	2713	2711	2724	2699	2660	2675	2649	2674	2667	2678	2626	2677	2680	2695	2766	2920	2948	2837	2949	3165	3277	3113	-5.0	0.7
BRICS	2673	2617	2556	2517	2525	2602	2638	2622	2635	2667	2744	2837	2950	3204	3436	3623	3784	3951	4052	4113	4425	4614	4652	4732	4787	4824	4732	4890	5081	5240	5219	-0.4	3.3
Europe	1128	1130	1134	1146	1167	1206	1258	1258	1226	1227	1238	1235	1245	1239	1228	1192	1170	1148	1157	1114	1124	1083	1080	1065	1044	1050	1044	1045	1039	1012	972	-4.0	-1.2
European Union	743	728	716	712	700	709	726	719	688	679	679	688	693	698	715	707	707	693	701	671	698	684	690	690	672	654	639	634	628	607	563	-7.1	-0.9
Belgium	13	13	13	12	12	12	12	13	13	14	14	13	13	14	14	14	14	14	14	15	16	16	14	15	13	11	15	15	12	16	14	-12.7	0.0
Czechia	41	39	37	36	34	33	33	34	32	29	31	32	32	34	35	33	34	34	33	32	32	33	33	31	30	29	27	28	28	27	23	-12.7	-1.4
France	112	119	120	126	124	128	132	129	125	127	131	132	134	136	137	137	136	134	136	129	136	137	135	137	138	139	132	129	135	132	120	-8.7	-0.4
Germany	186	168	163	152	145	145	143	143	136	137	135	135	134	136	138	138	141	139	136	129	132	125	126	124	121	120	115	115	112	103	96	-6.9	-1.7
Italy	26	26	28	29	30	30	31	31	31	29	28	27	29	30	29	31	30	31	33	32	33	32	35	37	37	36	34	34	35	34	35	3.1	1.1
Netherlands	61	67	67	69	67	67	75	67	64	60	58	61	61	59	68	62	61	60	68	63	71	66	66	70	60	48	46	42	36	33	28	-15.6	-3.6
Poland	104	100	98	97	97	99	99	100	87	84	79	80	80	80	79	79	78	72	71	67	67	68	72	71	68	69	67	65	63	60	55	-7.7	-1.8
Portugal	3	3	3	3	3	3	4	4	4	3	4	4	4	4	4	4	4	5	4	5	6	6	5	6	6	5	6	5	6	6	6	7.0	2.4
Romania	40	36	34	33	32	32	33	32	29	28	28	28	29	29	28	28	28	28	29	28	28	28	27	26	26	27	25	26	25	25	23	-8.2	-1.1
Spain	35	34	34	33	32	32	33	32	32	31	31	33	32	33	33	31	31	30	30	30	34	32	33	34	34	33	33	33	33	33	34	1.8	0.3
Sweden	30	31	29	29	31	32	32	32	34	33	30	34	32	31	34	35	33	33	33	30	33	33	36	35	34	34	35	36	36	37	33	-10.7	0.4
United Kingdom	208	215	215	223	245	258	269	268	272	282	272	262	259	247	226	205	187	176	167	158	148	130	117	110	108	118	120	120	123	123	117	-5.2	-4.1
Norway	119	131	147	155	171	185	208	213	207	210	228	227	236	235	229	224	216	215	222	217	207	199	204	195	197	208	210	216	207	196	210	7.3	-0.4
Turkey	25	25	26	26	26	26	27	28	29	27	26	25	24	24	24	24	26	27	29	30	32	31	31	29	29	32	36	37	41	46	42	-7.8	2.4
CIS	1662	1570	1456	1362	1252	1217	1207	1167	1176	1212	1261	1302	1353	1440	1510	1555	1603	1643	1680	1593	1697	1736	1758	1785	1756	1744	1784	1850	1925	1950	1816	-6.9	1.8
Kazakhstan	91	92	90	79	71	64	64	66	65	66	79	84	91	102	115	119	128	132	144	148	157	160	165	169	163	146	151	168	178	177	169	-4.1	3.9
Russia	1293	1218	1139	1057	991	967	965	933	940	962	978	1008	1046	1119	1172	1203	1227	1239	1254	1190	1279	1299	1316	1331	1319	1334	1374	1429	1484	1505	1402	-6.9	1.8
Ukraine	136	119	112	100	87	82	74	75	75	76	76	76	77	76	78	79	82	85	84	79	79	86	85	86	77	64	63	59	61	61	57	-6.7	-1.4
Uzbekistan	39	40	41	43	46	49	49	52	55	55	55	56	58	56	58	57	59	60	62	61	60	56	55	53	49	49	50	51	55	55	46	-15.8	-0.9
America	2543	2568	2589	2579	2684	2726	2807	2849	2880	2857	2887	2933	2916	2933	3001	3013	3076	3089	3120	3063	3132	3230	3282	3368	3527	3522	3399	3476	3638	3733	3547	-5.0	1.0
North America	1924	1924	1939	1910	1993	2007	2034	2038	2052	2032	2038	2063	2043	2023	2047	2034	2074	2091	2112	2080	2124	2201	2249	2327	2486	2501	2400	2508	2708	2832	2695	-4.9	1.4
Canada	276	288	297	319	341	352	361	367	368	367	375	379	387	389	401	402	419	422	409	394	399	414	428	449	470	473	480	509	530	527	504	-4.4	1.5
United States	1647	1637	1642	1592	1652	1655	1673	1671	1683	1665	1662	1684	1656	1635	1646	1632	1655	1670	1702	1686	1725	1787	1821	1879	2016	2027	1920	1998	2178	2305	2190	-5.0	1.4
Latin America	620	644	651	668	691	719	773	811	828	825	849	869	873	910	954	979	1002	998	1009	983	1007	1029	1033	1041	1041	1022	999	969	930	901	853	-5.4	0.0
Argentina	47	49	52	56	62	66	71	75	79	81	83	83	82	86	87	84	85	81	82	79	78	74	75	73	72	73	74	73	75	78	72	-7.0	-0.7
Brazil	105	105	105	106	111	112	119	127	134	142	148	152	168	179	183	195	207	217	229	232	248	250	251	253	267	278	285	294	297	312	324	3.8	4.0
Chile	8	8	9	8	8	8	8	8	8	8	9	9	9	9	9	9	10	9	10	10	9	10	13	15	12	12	13	13	14	13	13	-2.3	2.2
Colombia	48	46	48	49	50	59	65	67	73	76	72	72	68	73	75	79	84	87	93	98	106	121	125	130	131	129	125	124	120	120	104	-13.2	1.9
Mexico	196	204	203	207	206	204	214	226	226	225	229	234	239	258	264	264	262	251	237	224	223	224	219	216	209	190	180	165	158	148	147	-0.9	-2.2
Venezuela	145	161	162	171	178	192	217	229	227	210	219	228	210	198	219	224	223	218	223	204	201	203	199	199	191	184	169	149	115	82	51	-37.2	-7.0
Asia	1669	1714	1741	1790	1873	1994	2050	2071	2082	2105	2183	2252	2329	2505	2727	2918	3086	3229	3315	3467	3720	3898	3906	3986	4045	4038	3927	4029	4194	4358	4381	0.5	3.5
China	881	887	904	933	986	1065	1091	1081	1079	1074	1124	1176	1227	1375	1534	1671	1793	1917	1967	2052	2236	2386	2401	2467	2507	2516	2363	2455	2568	2686	2749	2.3	4.6
India	280	291	294	298	308	323	328	338	337	344	351	358	368	381	393	401	406	425	445	484	504	521	524	522	533	537	551	552	574	578	590	2.2	2.6
Indonesia	169	183	187	194	208	215	223	228	229	238	238	242	248	255	265	280	314	318	323	348	374	425	429	437	431	406	414	426	451	473	437	-7.7	3.1
Japan	76	80	81	89	91	99	103	107	110	105	107	105	100	87	98	104	105	95	93	99	103	55	32	32	30	34	37	43	52	53	50	-5.7	-3.7
Malaysia	48	52	53	55	57	64	70	74	74	73	77	78	81	85	92	96	93	93	96	90	89	87	87	92	94	96	97	96	99	97	89	-8.7	0.7

VARIABLES OF INTEREST

VARIABLE NAME	DESCRIPTION	DATA TYPE
Total Energy Production	Electricity production corresponds to gross production. It includes public production (production of private and public electricity utilities) and industrial producers for their own uses, by any type of power plant (including cogeneration).	Double
Total Energy Consumption	The total energy consumption is the balance of primary production, external trade, marine bunkers and stock changes.	Double
CO2 Emissions from Fuel Combustion	CO2 emissions cover only the emissions from fossil fuel combustion (coal, oil and gas).	Double
Refined Oil Products Production	Oil products are all liquid hydrocarbons, in particular, LPG production (Liquid Petroleum Gas) includes LPG	Double

Oil Products Domestic Consumption	from natural gas separation plants.	
Country	Contains 45 countries	Character
Year	Contains 3 decades	Integer

DATA PRE-PROCESSING

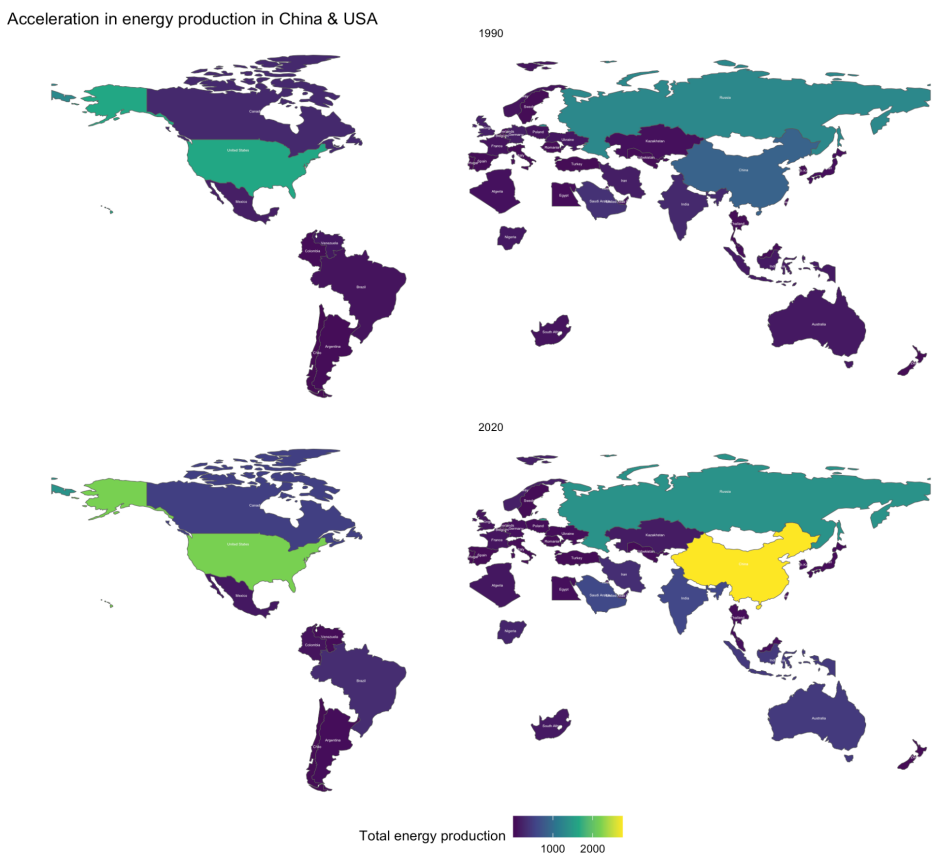
1. Since the data was present in multiple sheets, the data was merged into a single tibble using **map_dfr** function (used to iterate over sheets). A custom function named **read_sheet_to_tibble** was written to perform reading of data from excel sheet for desired range of cells and attaching the name of sheets (variables of interest) to the first column. This process made it easier to analyze the data as it brought everything together in one place.
2. Following functions were used to tidy the data:

- **Separate**: Separate `www.enerdata.net` column (contained data such as *Total energy production_Europe*) into category and country based on underscore separator.
- **Pivot_longer**: Reshape the data frame from wide to long format to have an year and a value column.
- **Pivot_wider**: Widen category column.

3. Used **geojson_read** function to read geojson data from the specified URL and store it as an *sp* object (spatial data frame). This data frame was merged to energy data frame to plot maps.

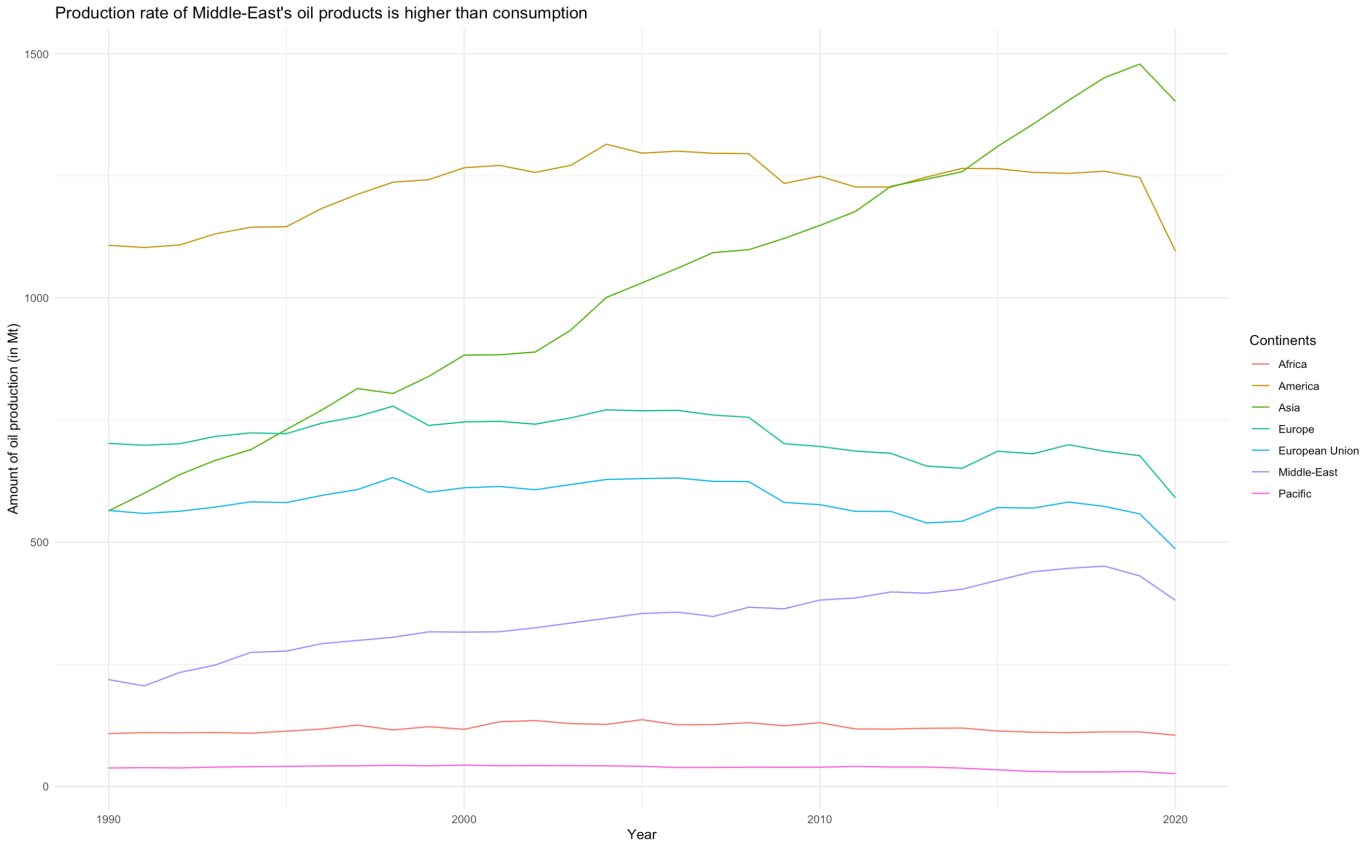
OBSERVATIONS AND TAKEAWAYS

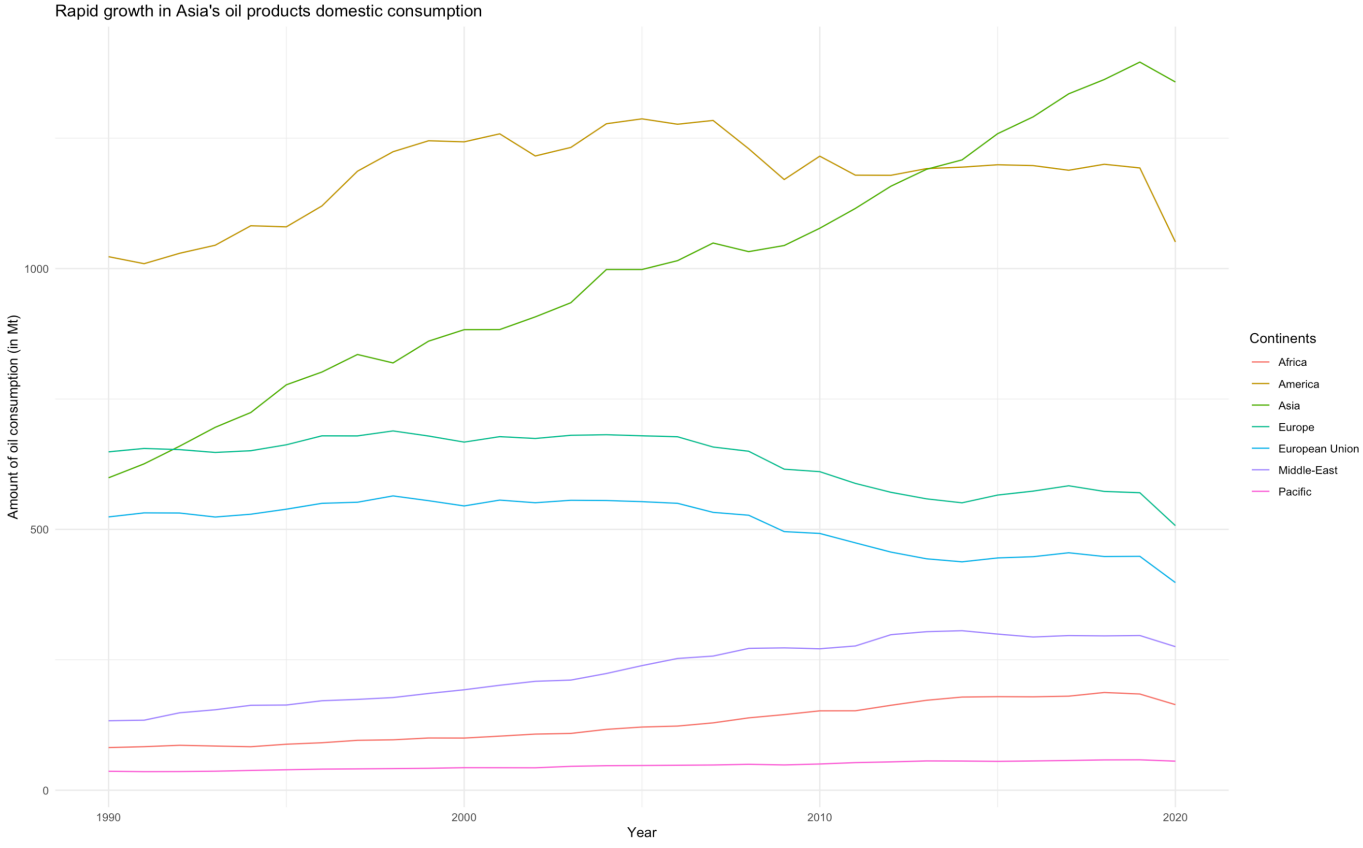
1. **Acceleration in Energy Production in China and USA over Three Decades:**
   - This acceleration reflects significant advancements in energy infrastructure, technological innovation, and resource extraction techniques.
   - Factors such as industrialization, urbanization, and economic growth have likely contributed to the heightened demand for energy production in these countries.
   - The observed trend underscores the pivotal role of China and the USA in global energy dynamics and their status as major players in the international energy market.



2. Production Rate of Middle-East's Oil Products Higher than Consumption:

- The Middle East region exhibits a notable disparity between the production and consumption rates of oil products.
- With abundant oil reserves and significant refining capacities, countries in the Middle East have been able to maintain high levels of oil product production.
- However, domestic consumption within the region does not match the production rates, leading to surplus oil products available for export to global markets.
- This surplus production capacity positions the Middle East as a key supplier of oil products to international markets, contributing to its strategic importance in the global energy trade landscape.


Production rate of Middle-East's oil products is higher than consumption
Amount of oil production (in Mt)
1500
1000
500
0
1990
2000
2010
2020
Year
Continents
Africa
America
Asia
Europe
European Union
Middle-East
Pacific


Rapid growth in Asia's oil products domestic consumption
Amount of oil consumption (in Mt)
1000
500
0
1990
2000
2010
2020
Year
Continents
Africa
America
Asia
Europe
European Union
Middle-East
Pacific

Reply

Attach

Cancel Post Reply

Vandit Gupta (https://northeastern.instructure.com/courses/170748/users/265163)

Feb 13, 2024

Exploring the Intersections of Literature: A Dive into Goodreads' Book Data by Vandit Gupta 1. Da

Exploring the Intersections of Literature: A Dive into Goodreads' Book Data

by Vandit Gupta

1. Dataset Description:

For my homework 2, I explored the **Goodreads-books** dataset, provided by **Jealousleopard** on Kaggle in 2020 **(Jealousleopard. (2020). Goodreadsbooks [Data Set]. Kaggle. https://www.kaggle.com/datasets/jealousleopard/goodreadsbooks/ (https://www.kaggle.com/datasets/jealousleopard/goodreadsbooks/))**. This dataset gathers extensive information from Goodreads, a site beloved by readers for book reviews and recommendations. I was attracted to this dataset for its potential to shed light on various aspects of book popularity and reader engagement. My exploration was driven by specific questions regarding the characteristics that contribute to a book's reception. I aimed to explore the distribution of average ratings to see how books are generally appreciated by readers, investigate the distribution of the number of pages to understand if book-length influences reader preferences, identify the top authors in terms of the number of books written to gauge author productivity and popularity and determine the top publishers based on the number of books published to understand their impact on the literary market.

2. Dataset Structure

The Goodreads-books dataset is a rich repository that captures a diverse array of information about

45,641 books listed on Goodreads. It is structured with various fields, each offering unique insights into the literary world:

• **bookID:** A unique numeric identifier for each book, useful for database management and referencing specific books within the dataset.

• **title:** The full title of the book, which provides insight into the book's content and theme.

• **authors:** The name(s) of the author(s) who wrote the book. In cases where there are multiple authors, their names are combined in a single string separated by commas.

• **average_rating**: A numeric value representing the average rating given to the book by Goodreads users, on a scale ranging from 1 to 5. This offers a general idea of the book's reception among readers.

• **isbn** and **isbn13**: These are unique identifiers for books, with ISBN referring to the 10-digit International Standard Book Number and ISBN13 the 13-digit version. These numbers are crucial for book identification and inventory management.

• **language_code:** A code indicating the language in which the book is written. This information is vital for readers looking for books in specific languages.

• **num_pages:** The total number of pages in the book, giving an idea of its length.

• **ratings_count:** The total number of ratings the book has received on Goodreads, which can indicate its popularity and how widely it has been read.

• **text_reviews_count:** The number of written reviews the book has received, offering insight into how many readers have engaged with it deeply enough to leave feedback.

• **publication_date**: The date when the book was published, providing context for its historical and cultural background.

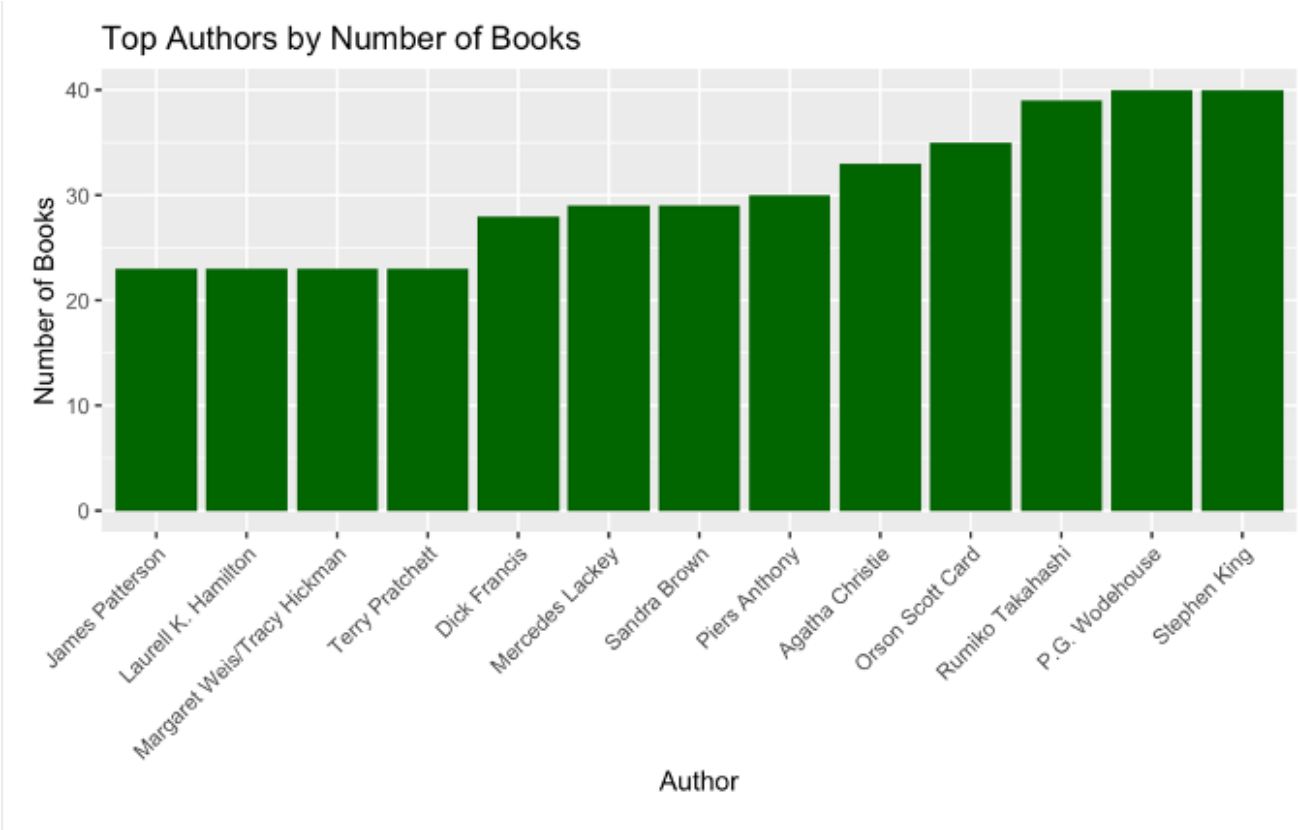
• **publisher**: The company or individual responsible for publishing the book, which can affect its distribution, marketing, and availability.

3. Data Preprocessing Steps

To prepare the Goodreads-books dataset for analysis, I performed several essential preprocessing steps. First, I removed rows with missing values to ensure the dataset's completeness. Next, I corrected the data types for various columns, such as converting **bookID** to integer and **average_rating** to numeric, ensuring they accurately reflected the nature of the data. I also verified the absence of duplicate entries, maintaining the dataset's uniqueness. Lastly, I examined numeric columns like **num_pages**, **average_rating**, **ratings_count**, and **text_reviews_count** for outliers using boxplots, identifying and addressing any anomalies to ensure a clean and reliable dataset for

analysis.

4. Visualization and Analysis:



In examining the Goodreads-books dataset, a standout visualization is the bar chart of the most published authors. It shows a range of prolific writers, with **Stephen King, P.G. Wodehouse, and Rumiko Takahashi** leading the pack. The chart not only reflects their productivity but also the diverse genres they represent, suggesting a varied reader interest in Goodreads. This insight is invaluable as it highlights the importance of an author's output volume and the potential influence on their popularity and readership engagement.

In conclusion, the visual exploration of the Goodreads-books dataset not only validates the prolific nature of certain authors but also underscores the diverse literary tastes among Goodreads users. This analysis paves the way for more nuanced inquiries into the relationship between a book's attributes and its success, further enhancing our understanding of the publishing landscape and reader preferences.

Edited by **Vandit Gupta (https://northeastern.instructure.com/courses/170748/users/265163)** on Feb 13 at 9:46pm

Reply

Attach

Cancel Post Reply

Harikrishnan Unnikrishna Pillai (https://northeastern.instructure.com/courses/170748/users/261004)

Feb 13, 2024

The dataset I chose gives information regarding incidents that took place across the world giving a

The dataset I chose gives information regarding incidents that took place across the world giving a record of the deaths of migrants who have died at the external borders of states, or in the process of migration towards an international destination, regardless of their legal status. The **source for the dataset** is https://missingmigrants.iom.int/ (https://missingmigrants.iom.int/) . As an Indian who has a lot of family members who have migrated to various parts of the world and myself being born in the Middle East, I was curious to dig deep into the dataset when I stumbled upon it. I chose this dataset to find out any patterns behind migrant deaths and to try and figure out what the possible reason could be for the same.

The dataset was pretty structured and was downloadable in a tabular format. The **variables of interest** are:

Year - Year when the incident was recorded

Total_Dead_Missing - Sum of people who passed away or are missing, and presumed dead

Origin_Country - Nationality of the deceased

Origin_Region - Region of the origin country of the deceased

Death_Cause - Cause of death

Incident_Country - Country where the deaths were reported

Migration_Route - The route taken by the deceased for migration

The preprocessing steps that was undertaken are:

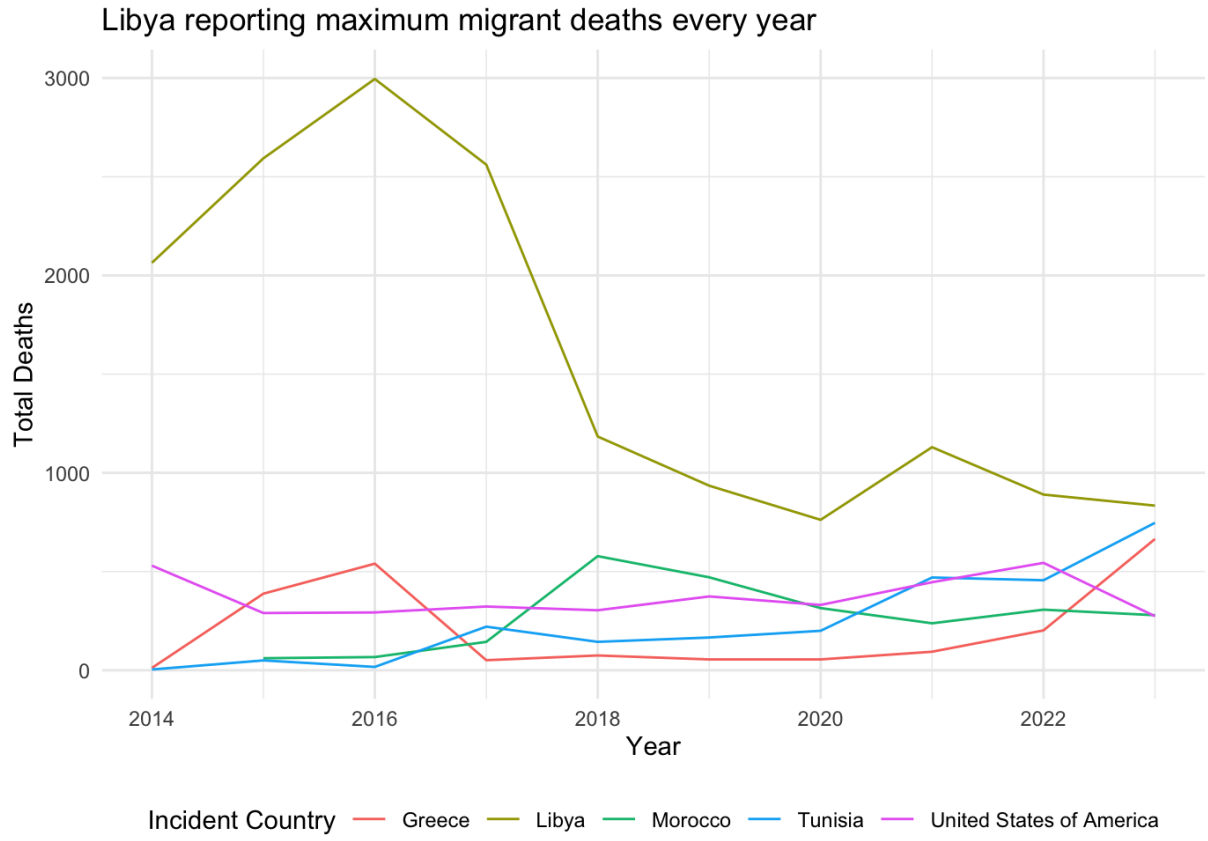
- Removed unwanted columns not relevant to the data analysis

- Filtering based on the Source Quality of the observational units (minimum 3 out of 5) to maximize authenticity

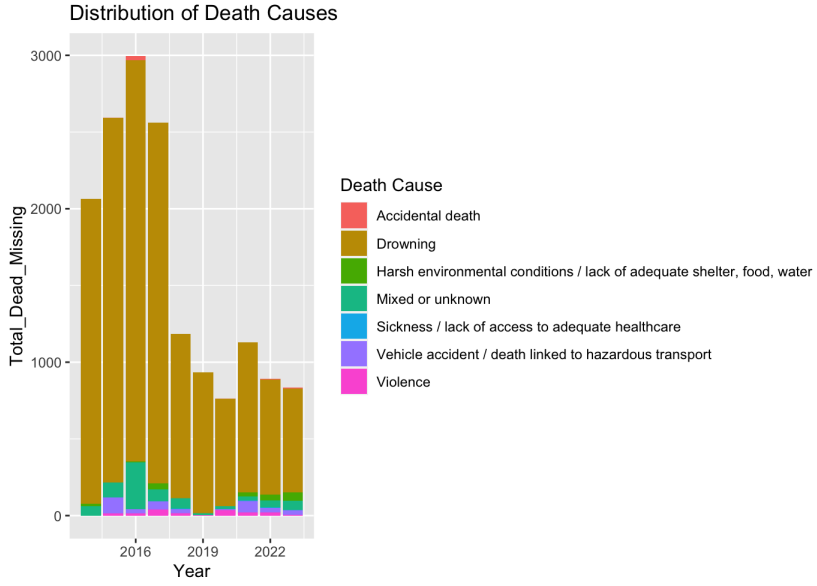
- Imputed data with missing values for Missing Values for Total Number of Dead and Missing

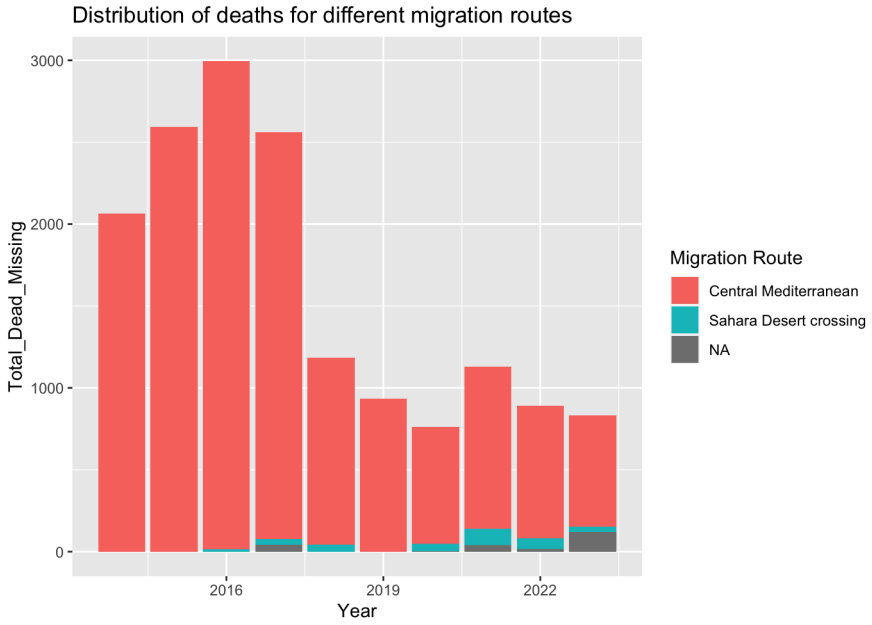
- Filtered out 2024 data since there is not enough data for the current year yet to draw any conclusion

Data analysis:



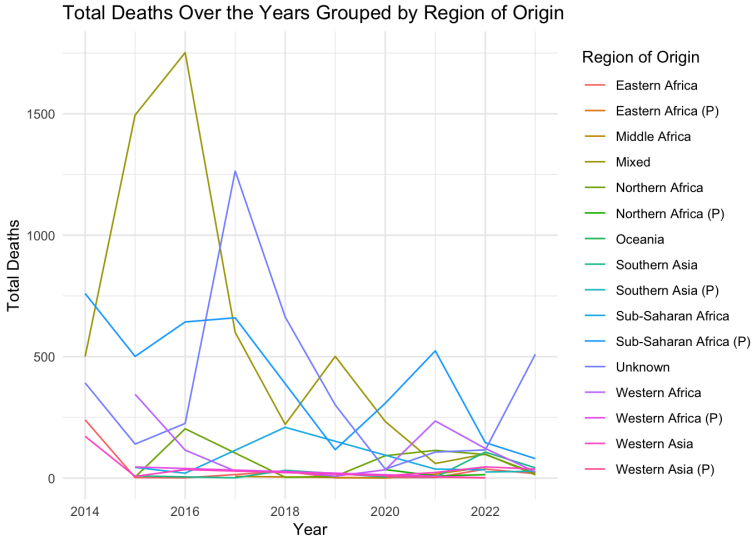
From the visualization, it can be seen that Libya is the country where the highest number of incidents and deaths have been reported consistently from 2014 up until last year.





Looking into the causes of the deaths in Libya, it was observed that the deaths were predominantly due to drowning in the Mediterranean sea.

I looked into the data to see the nationalities of the deceased and found that the origin of most of these people were from the African continent itself, which implied that these were people trying to migrate from their homeland across the sea(to Europe) in pursuit of a better livelihood.



To conclude, the dataset shows us that most migrant deaths have been reported in Libya for the last 10 years. The majority of the deaths were of citizens of several neighboring countries which indicates people originating from the continent were possibly trying to migrate to Europe in pursuit of better livelihood and resources for survival. Libya's geographic position and the distance to the European coast as compared to its neighboring coastline countries (termed as Northern African Rim) may be the reason for their decision to migrate. The harsh Mediterranean route has unfortunately proven to be fatal for these migrants which can be seen from the obvious trend from the initial graph showing Libya's reported death count of migrants.

To understand the geographic placement of the countries, I had also referenced **https://www.worldatlas.com/geography/mediterranean-countries.html (https://www.worldatlas.com/**

geography/mediterranean-countries.html)

Reply

Attach

Cancel Post Reply

-

Anikesh Ganesh Kamath (https://northeastern.instructure.com/courses/170748/users/229898)

Feb 14, 2024

Flash Paper - Formula 1 Lap Data Analysis Dataset Description For this problem I have chosen a K

Flash Paper - Formula 1 Lap Data Analysis

Dataset Description

For this problem I have chosen a Kaggle dataset which contains the Formula 1 races, drivers, constructors, qualifying, circuits, lap times, pit stops, championships from 1950 till the latest 2023 season.

Link: https://www.kaggle.com/datasets/rohanrao/formula-1-world-championship-1950-2020 (https://www.kaggle.com/datasets/rohanrao/formula-1-world-championship-1950-2020)

(https://www.kaggle.com/datasets/rohanrao/formula-1-world-championship-1950-2020) For reference, here is some background information about Formula 1:

- Formula 1 is an extremely competitive racing sports which takes place across the world.
- In Formula 1 there are 10 teams (called constructors), and each each team has 2 drivers (for a total of 20 drivers).
- Races take place all in circuits across the world.
- A race consists of 40-70 laps which usually take between 2 to 3 hours.
- One day before the race takes place there is a “qualifying” session where the order for the grid is determined.

- One championship lasts one calendar year.

Formula 1 interests me because it combines the glitz and glam of a high performance sport with extremely complex technical specifications and machines. Each component is a Formula 1 car is tuned to maximize the car's performance. There are thousands of points of data that is fed from the car to the team's headquarters every second during the race. There are dedicated teams who analyze this data on the go, compare it other team's performance, and make strategy calls.

This dataset in particular interests me because it has a lot of important data points about lap times from the early 2000s, and allows for a comparison between the older cars vs the current generation cars, and also between the different drivers. I would like to answer the following questions:

1. How do race lap times compare between the early 2000s vs today?
2. Has a driver's performance improved, reduced or remained consistent throughout this time, especially with the different regulation changes taking place?
3. How a driver perform compare to his teammate in the same car? How does a driver perform compared to other drivers with similar capabilities?

Structure of Dataset

From the different CSVs available in the dataset, I have chosen two which are required for my analysis. I have described the data below with the variables of interest.

lap_times.csv

Lap times for races

Variable	**Description**	**Type and Examples (using the str() command in R)**	**Variable of Interest Rationale**
raceId	ID of the race	num; 841, 842	To select specific races to compare
driverId	ID of the driver	num; 20, 21, 22	To select specific drivers to compare
lap	Lap number	num; 1, 2, 3	To compare times over different laps
time	Time in mm:SS.SSS	'hms' num; 1:28:923, 1:38:094	Lap time is the main comparison variable

races.csv

Races

Variable	**Description**	**Type and Examples (using the str()**	**Variable of Interest Rationale**

		command in R)	
raceId	ID of the race	num; 841, 842	To select specific races to compare
year	Year of the race	num; 2007, 2008	To compare times over different years
circuitId	Circuit ID of the race	num, 1, 2, 3	To select specific circuits to compare

I also referenced **drivers.csv** and **circuits.csv** to find the specific IDs I required for the analysis.

Preprocessing Steps

For preprocessing the data, I need to selected specific drivers and races. I wanted to analyze the lap times for the Belgium Grand Prix, and specifically I wanted to analyze the lap times of the drivers Lewis Hamilton, Valtteri Bottas, and Max Verstappen. The **lap_times.csv** contains the lap times, whereas the **races.csv** contains the circuit information. I needed to choose specific columns from the dataset and then join the two data frames to enable me to create the required visualizations. I have described the preprocessing steps below.

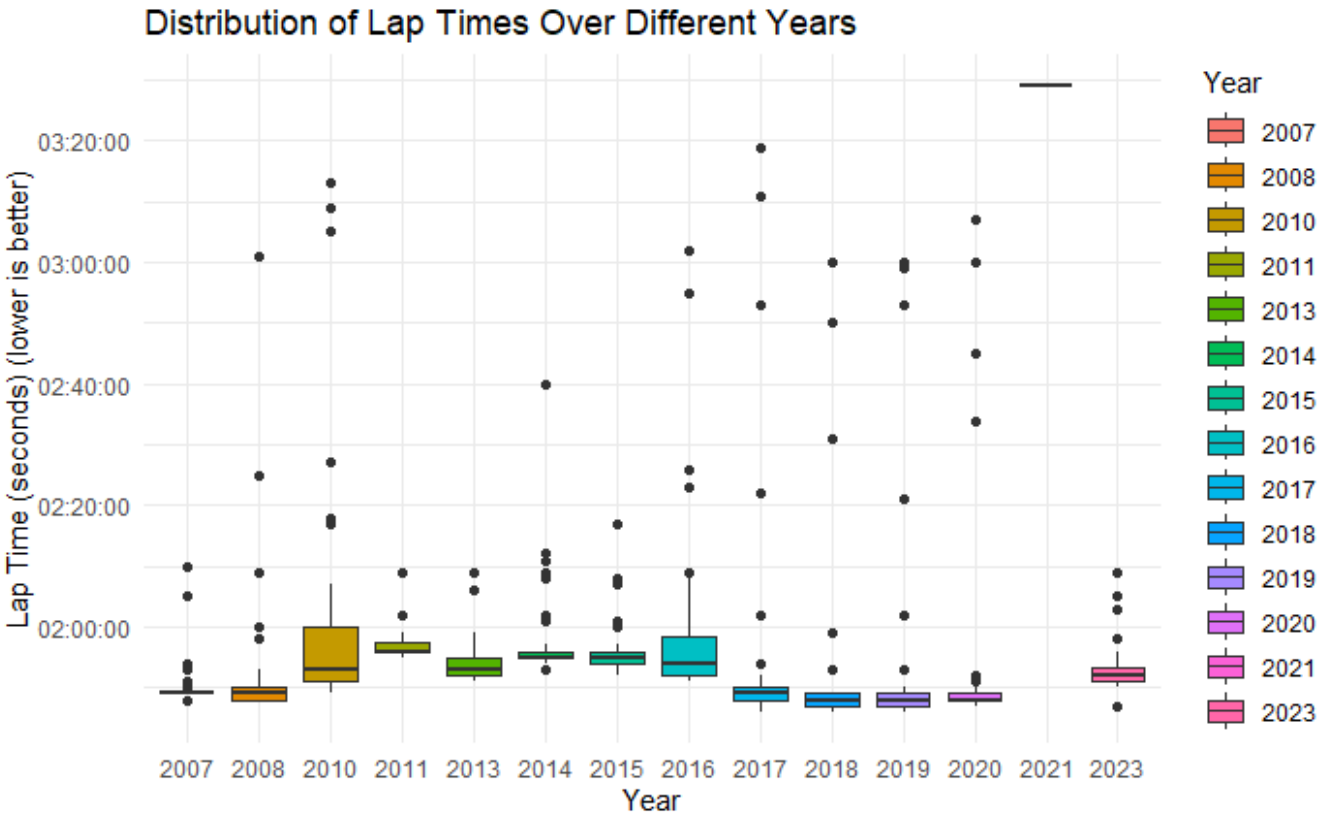
1. Loaded the CSVs into R Studio and loaded the required libraries (tidyverse, tidyr, readr, dplyr).
2. Selected only lap times from lap_times.csv with *driverId* as 1 (Lewis Hamilton) and assigned to a data frame called **hamilton**.
3. Selected only the races from races.csv with *circuitId* as 13 (Belgium Grand Prix) and assigned to a data frame called **races**.
4. Joined **races** with **hamilton**, and discarded all lap times that do not have a *year* value.
5. Selected only lap times from lap_times.csv with *driverId* as 1 (Lewis Hamilton), 822 (Valtteri Bottas), and 830 (Max Verstappen) and assigned to a data frame called **combined**.
6. Joined **races** with **combined**, discarded lap times that do not have a *year* value, and selected lap times with the *year* as 2020 and *lap* number greater than 15.
7. Created a new column in **combined** called *driver* to store the driver name as a string.

Key Observations

I have included two key observations below.

Box Plot Distribution for Lap Times over Different Years

This box plot shows the distribution of lap times from 2007 to 2023 for Lewis Hamilton in the Belgium Grand Prix. The lap times are grouped by year.



Conclusions:

- The mean lap times have generally reduced over the years, indicated by the mean reducing over the years.
- Lewis performed very consistently from 2017 to 2020, indicated by the similar box plot for these years.
- There was dip in performance in 2023, indicated by the higher box plot in 2023 compared to 2020.
- The dip in performance could have multiple possible reasons - few of them could be that in 2021 Mercedes' revolutionary dual-axis steering (DAS) system was banned, and in 2022 a lot of regulations were changed related to aerodynamics. This meant the 2023 car was drastically different in terms of a complete performance package compared to the 2020 car.

Notes:

- In 2021, the race was called off after 1 lap due unfavorable race conditions.
- In 2022, Lewis Hamilton crashed before completing the first lap and hence there is no lap time data.

References:

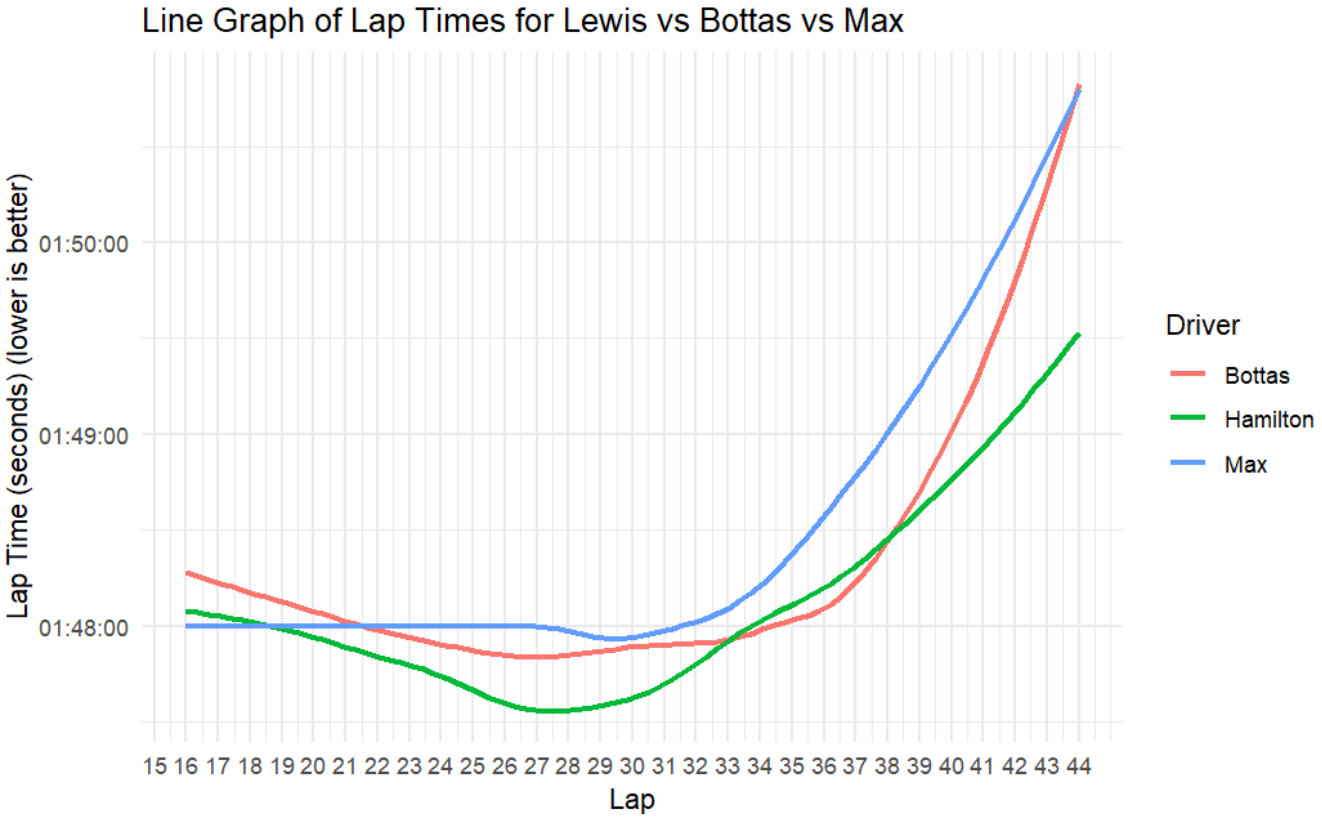
- **https://en.wikipedia.org/wiki/2021_Belgian_Grand_Prix (https://en.wikipedia.org/wiki/2021_Belgian_Grand_Prix)**
- **https://en.wikipedia.org/wiki/2022_Belgian_Grand_Prix (https://en.wikipedia.org/wiki/**

2022_Belgian_Grand_Prix)
- https://en.wikipedia.org/wiki/History_of_Formula_One_regulations (https://en.wikipedia.org/wiki/History_of_Formula_One_regulations)

Line Graph of Lap Times for Lewis vs Bottas vs Max

For the final comparison, we will examine the lap times for three different drivers in the same race - 2020 Belgium Grand Prix.

Valtteri Bottas was Lewis Hamilton's team mate in 2020, and drove a similar car from Mercedes. Max Verstappen was part of another team called Red Bull Racing, and had either the 2nd or 3rd best car.



Conclusions:

- As we can see the Mercedes drivers consistently outperform Max in the Red Bull throughout the race.
- Another observation is that even though Bottas is in a similar car to Hamilton, Hamilton is still able to outperform him in most parts of the race.

Edited by Anikesh Ganesh Kamath (https://northeastern.instructure.com/courses/170748/users/229898) on Feb 14 at 9:42am

Reply

Attach

Cancel Post Reply

-

Melissa Rejuan (https://northeastern.instructure.com/courses/170748/users/164393)

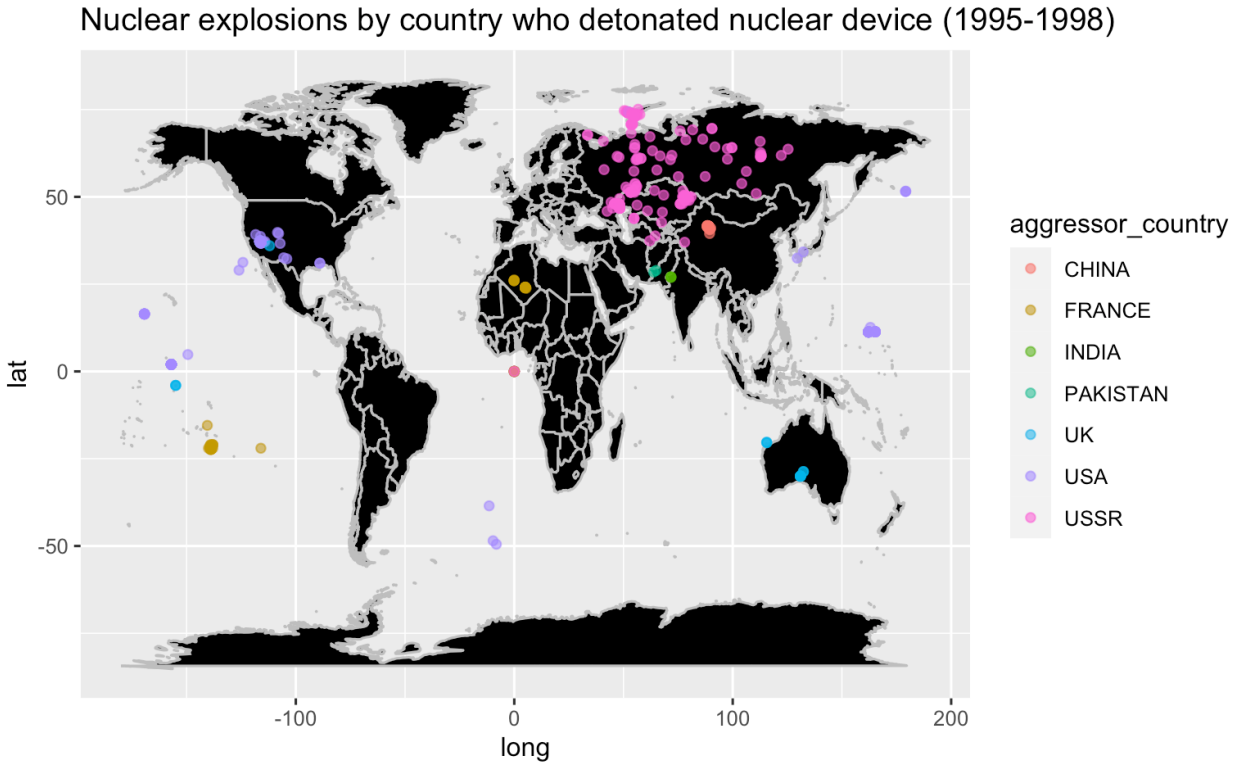
Feb 14, 2024

The data I used was nuclear explosion data from 1995-1998. I found this on the Data is Plural GitH

The data I used was nuclear explosion data (https://github.com/data-is-plural/nuclear-explosions/blob/master/data/sipri-report-explosions.csv) from 1995-1998. I found this on the Data is Plural GitHub page. The data was extracted from the Stockholm International Peace Research Institute (SIPRI), where information about nuclear explosions conducted by various countries was reported. I chose this dataset because the comeback of nuclear energy has been a hot topic globally as we are in an energy crisis. However, there have been many incidents of nuclear disasters throughout history, which makes bringing back nuclear energy questionable as they have detrimental health effects. Therefore, in the dataset, I wanted to explore how much nuclear energy has been used in the past and take a look at those explosions and their effects. Some questions I wanted answered were ‘how often was nuclear power used?’, ‘where was it used most often’, ‘who were using it the most’? Exploration of these questions would allow for further investigation through research.

The dataset contains information about nuclear explosions such as the name of them, where (region and coordinates), who detonated the nuclear devices, measurements of the impact of the explosion, purpose of the explosion, etc. For the purposes of my interest in how I want to use the data and making the dataset tidy, I removed columns that were not useful for the exploration I had in mind or were not clear enough to use, such as the purpose of explosion. Each value was an acronym in the purpose of explosion column, but there was no explanation of that column I could find that would identify what the acronyms were. Therefore, I removed it since I didn’t have a full understanding of it. The columns I did choose included ‘id_no’, ‘country’, ‘lat’, ‘long’, and ‘year’. These columns can help explore trends throughout time and provide geographical insights. In addition, I needed to transform some data to add more clarity to the dataset. I renamed the 'country' column to 'aggressor_country' because the original name is misleading. The country is not where the nuclear explosion happened, but it is who detonated the nuclear device. I also renamed "PAKIS" to "PAKISTAN" for clearer

labeling. Then, I removed any null values. This is especially important for longitude and latitude because when plotting coordinates, I don't want there to be missing values to cause an error. Lastly, I grouped years and countries by count to explore the data and to see if there anything worth investigating with visuals. I saw that certain countries had detonated a significantly larger number of nuclear explosions than others, therefore I decided to create a map visualization to show this.



One of the visualizations I created was a map of nuclear explosions with each point labelled with the country who detonated the nuclear device. The map shows some interesting insights. In the years between 1995-1998, it seems that the USA and USSR were the top countries to detonate nuclear devices, indicating they were the top players in the era of nuclear power. In addition, the number of points is also quite a shocking number to see, as it shows how heavily nuclear power was used. Because of the vast number of points, I had to increase the transparency of the points to view it more clearly. Many of the nuclear explosions were also mainly in the US and Russia. This is useful information as these are focus areas to investigate the health impacts of nuclear explosions to individuals in later years.

Reply

Attach

Cancel Post Reply

Pavan R Rao (*He/Him*) (https://northeastern.instructure.com/courses/170748/users/265874)

Feb 14, 2024

Flash Paper - World CO2 Emissions Dataset Description The dataset being used in this flash paper

Flash Paper - World CO_2 Emissions

Dataset Description

The dataset being used in this flash paper is the historical GHG (green house gases) emissions, specifically CO_2, in the word from the year 1990 to 2020.

Carbon dioxide (CO_2) emissions are those stemming from the burning of fossil fuels and the manufacture of cement.
They include carbon dioxide produced during consumption of solid, liquid, and gas fuels and gas flaring.

Data for carbon dioxide emissions include gases from the burning of fossil fuels and cement manufacture, but excludes emissions from land use such as deforestation.
The unit of measurement is **kt** (kiloton). Carbon dioxide emissions are often calculated and reported as elemental carbon.
The were converted to actual carbon dioxide mass by multiplying them by 3.667 (the ratio of the mass of carbon to that of carbon dioxide).

Source:
For this, the dataset that I have chosen is the **"CO2 emissions (kt)"** from **The World Bank Data Set** and the **Climate Watch Historical GHG Emissions** (1990-2020). *2023. Washington, DC: World Resources Institute.*
This data set is available at The World Bank website - https://data.worldbank.org/indicator/EN.ATM.CO2E.KT (https://data.worldbank.org/indicator/EN.ATM.CO2E.KT) - and at - https://climatewatchdata.org/ghg-emissions (https://climatewatchdata.org/ghg-emissions)
License : Attribution-NonCommercial 4.0 International (CC BY-NC 4.0)

Reasoning for choosing the data set

CO_2 is the principal anthropogenic greenhouse gas that affects the Earth's radiative balance.
It is the reference gas against which other greenhouse gases are measured, thus having a Global Warming Potential of 1.
Burning of carbon-based fuels since the industrial revolution has rapidly increased concentrations of atmospheric carbon dioxide, increasing the rate of global warming and causing anthropogenic climate change.
It is also a major source of ocean acidification since it dissolves in water to form carbonic acid.

I am passionate about reducing my own carbon footprint, as well as help others contribute to doing the same so that the climate bounces back from it's already catastrophically high global warming. One way to help others is to show statistically significant data, and in order to do that, I wanted to explore the data around historic CO_2 emissions, especially from a trusted source such as the World Development Indicators by The World Bank.

The "burning" questions in my mind were:
1. What was total CO_2 emissions over the years (1990-2020) for all countries as well as clusters of countries based on income group and overall development?
2. Who were the top 20 polluters in the world historically?

Dataset Structure

The dataset used has the following structure:

- `Country Code`: The country code (example "USA" for United States).
- `Indicator Name`: The name of the gas indicator (here, it's only CO_2 emissions in kilotons).
- `Indicator Code`: The code of the gas indicator (here, it's only the code for CO_2, example - `"EN.ATM.CO2E.KT"`).
- `1960` to `2022`: The CO_2 emissions (in kilotons) for each year from 1990 to 2022.

Of particular interest for my analysis are the `"Country Name"` and the columns from `"1960"` to `"2020"` (which are the CO_2 emissions for that year for every country/country group).

Pre-processing and Tidying

Note: The dataset is supposed to contain the CO_2 emissions (kt) for each country from *1960* to *2022*. However, the dataset doesn't actually contain the emission values for many years.

The steps required for pre-processing and tidying this data to be able to perform some initial visual analysis to answer my questions are:

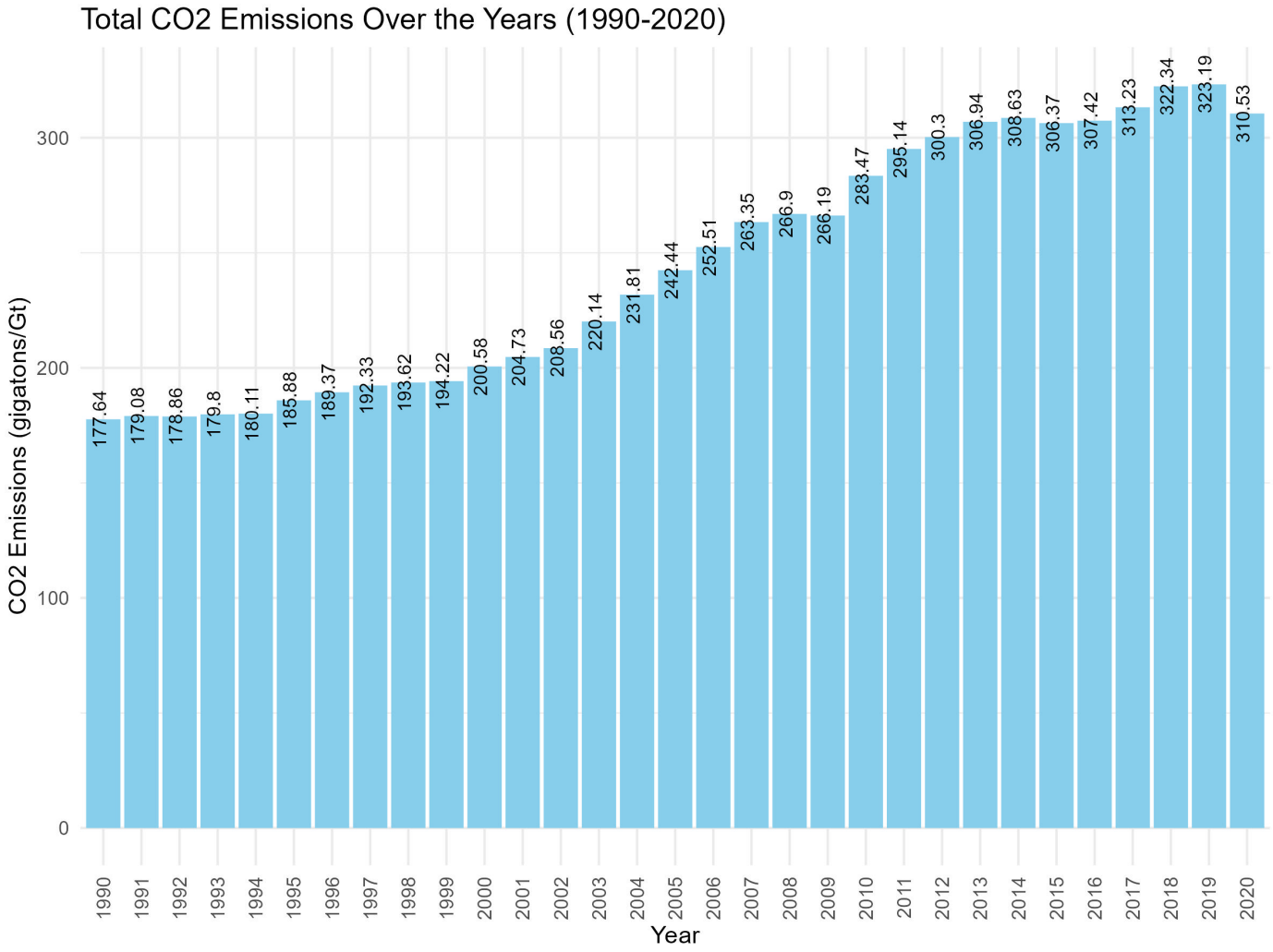
1. After obtaining the data from the source, load it into the R Markdown file where the analysis would be performed.
2. Select only the columns that contain the emission values for each year from *1990* to *2020* (omit *1960-1989*, *2021*, *2022*).

3. Pivot the data from wide to long format, so that there exists a single column for the year and another column for the emission values.
4. For *question 1*, aggregate the data of CO_2 emissions of all countries and country groups for each year.
5. This is then visualized in a bar plot to see the trend.
6. For *question 2*, first group the rows of the original data frame by country name so that the aggregated CO_2 emissions can be compared.
7. After that, aggregate (sum) the emissions for each country, and filter out the first 20 values so that we obtain an indication of the top 20 polluters.
8. This is then visualized using an overlapped line plot for analysis.

Interesting Figures And Analysis

Total CO_2 Emissions Over The Years (1990-2020)

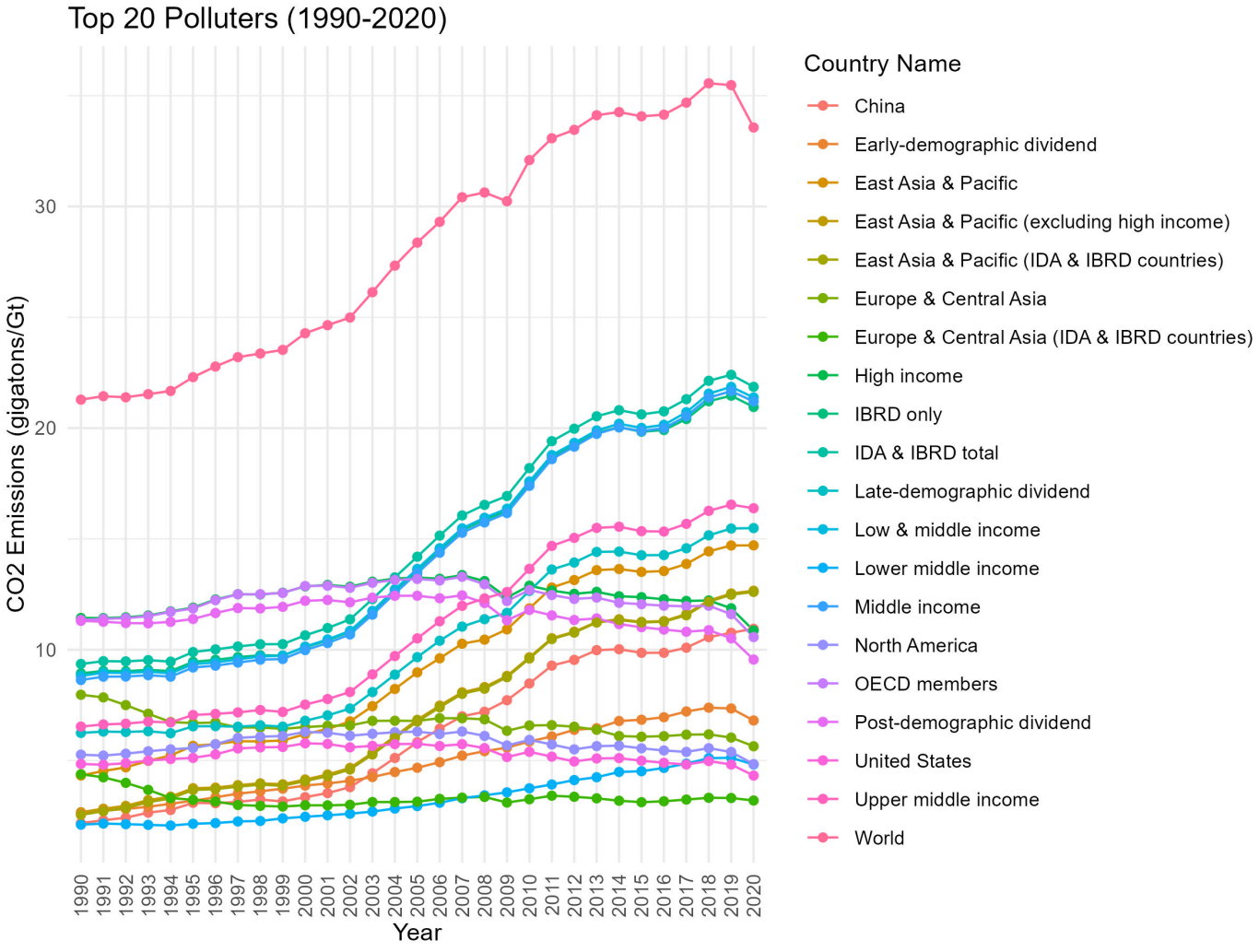
After aggregating the CO_2 emissions over the years for each country, the following bar plot is visualized so that the trend over the three decades is observed:



The bar plot shows the total CO_2 emissions over the years (1990-2020) for all countries. We can see that the CO_2 emissions have been increasing over the years and but has been decreasing in the recent years (2015-2020). This is a good trend for the world as a whole, as it shows that the world is moving towards reducing the CO_2 emissions.

CO_2 Emissions For The Top 20 Countries (1990-2020)

On grouping the emissions by country, aggregating the emissions and then filtering the top 20 countries in the emissions scale, we can visualise the data in the following line plot:

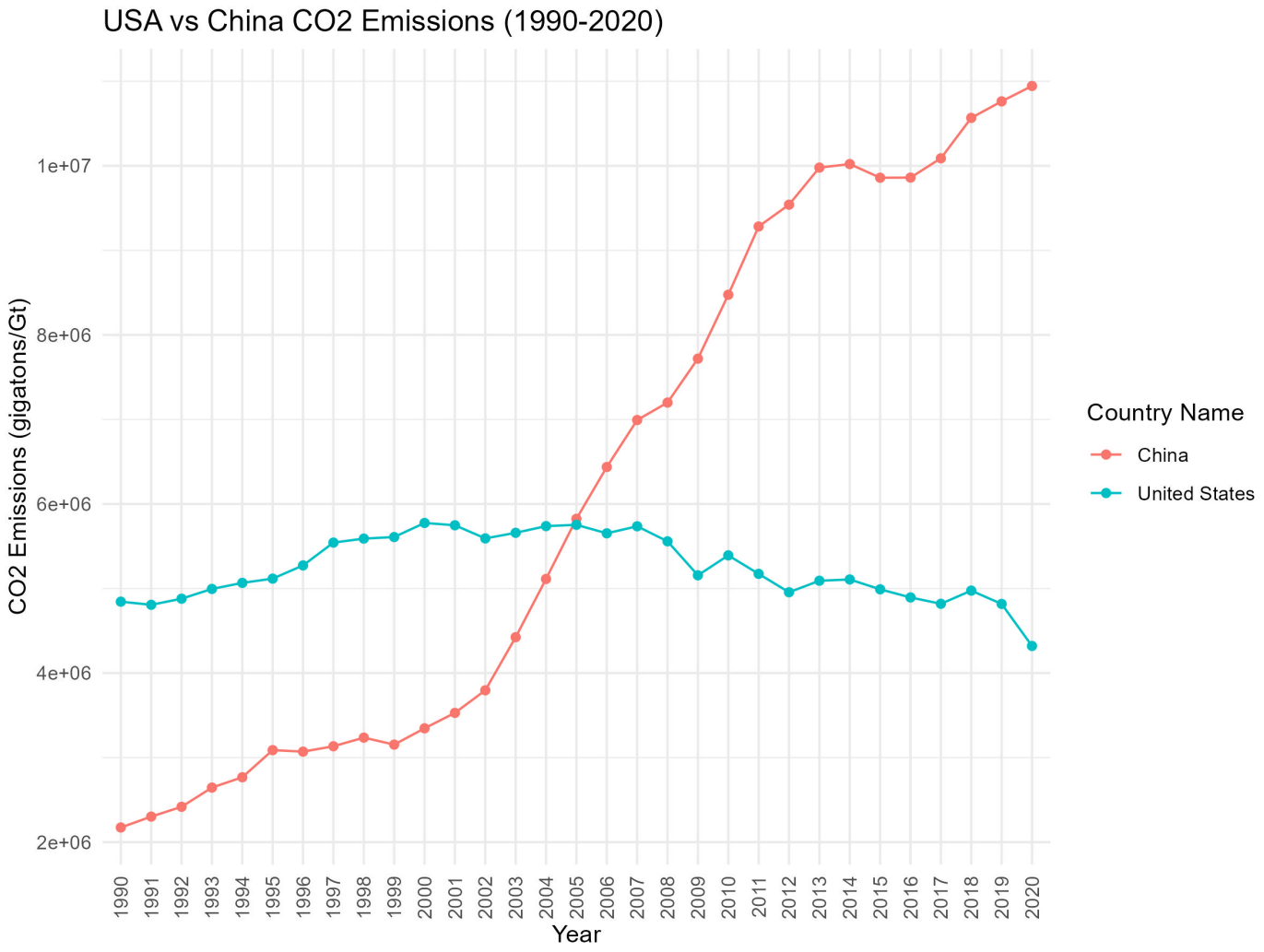


The line plot shows the CO_2 emissions over the years for the top 20 countries.
We can see that the CO_2 emissions have been increasing over the years for most of the countries.

We can observe that only China and the USA are in the top 20 polluters list (assuming we are including country groups).

The largest polluter has historically been the United States, but the emissions have been decreasing in the recent decade,
and China has been rising steadily. We can see the following plot to see the point where China

overtook USA in
CO_2 emissions.



We can observe that China overtook USA in global CO_2 emissions in 2005.

The takeaway from this, considering other world events around that time frame is that this rapid increase in the CO_2 emissions of
China coincided with their rapid increase in manufacturing and according to this data set, doesn't show signs of slowing down.

Another observation we can make from this is that the steady decline in the CO_2 emissions from around this same time could be attributed
to a related factor - the outsourcing of manufacturing, especially to China (although we would need to investigate more
data sets to further analyse and conclude on this particular observation).

Reply

Attach

Cancel Post Reply

Jie Li (https://northeastern.instructure.com/courses/170748/users/176283)

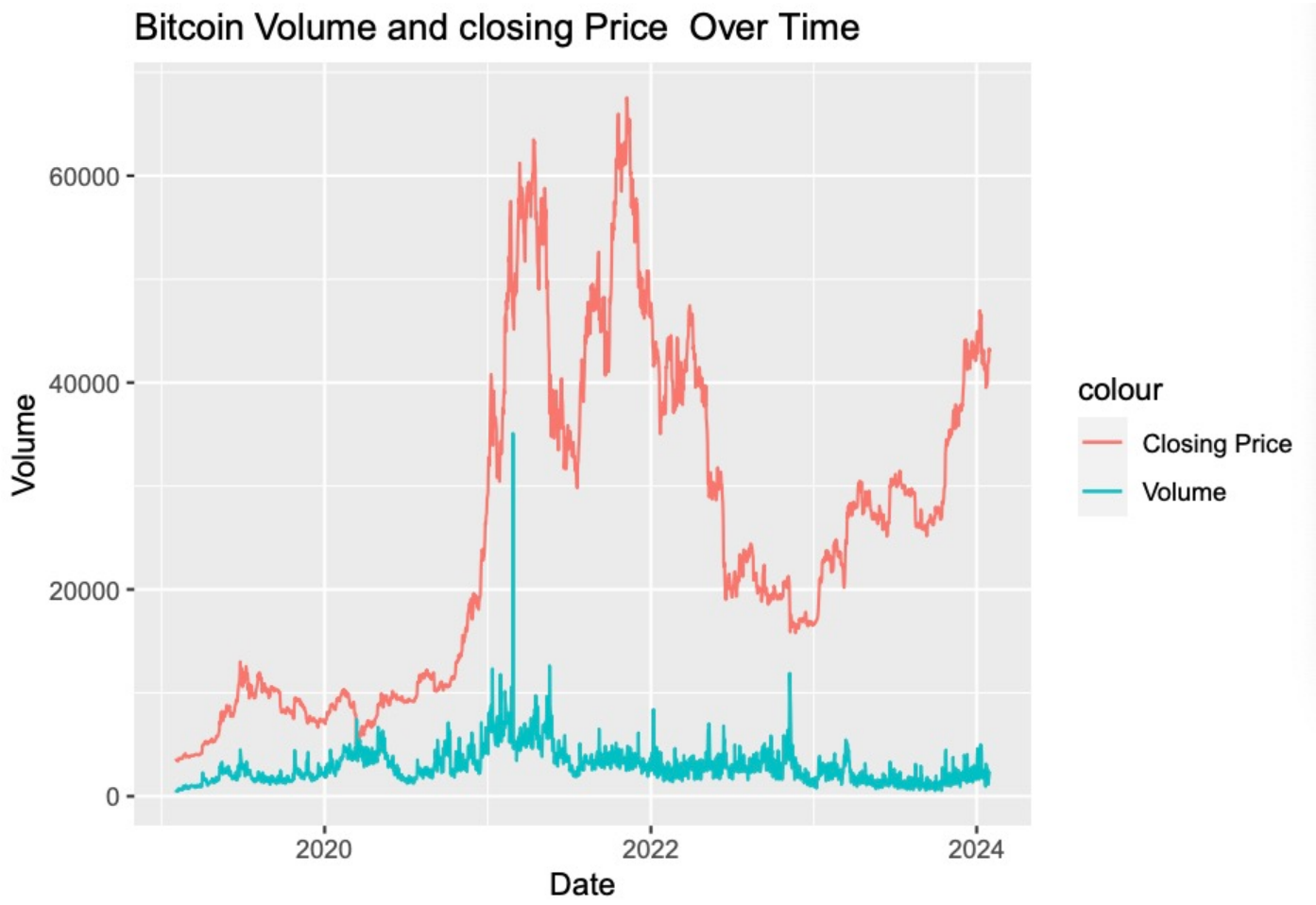
Feb 14, 2024

In homework 2 part A, I selected historical Bitcoin price data spanning the past five years, encompa

In homework 2 part A, I selected historical Bitcoin price data spanning the past five years, encompassing variables such as Date, Open Price, High, Low, Close, Adjusted Close, and Volume. The dataset was sourced from Yahoo Finance, accessible through the following link: https://finance.yahoo.com/quote/BTC-USD/ (https://finance.yahoo.com/quote/BTC-USD/) . As an investor myself, I was intrigued to explore potential trends or relationships within Bitcoin's price dynamics.

Upon acquiring the dataset, I engaged in preprocessing steps to ensure its suitability for analysis. Specifically, I addressed any null values by replacing them with "NA" and curated the variables of interest while also renaming columns for enhanced clarity.

In my visualization, I focused on examining the relationship between Bitcoin's closing price and trading volume. My takeaway from my visualization is that there seems to be a relationship between volume and price movement. During periods of sharp price increase or decrease, there is often a corresponding increase in volume. This is common in financial markets as higher volume can indicate more interest and activity, which often accompanies large price movements. - There might be a correlation between volume spikes and subsequent price movements. For instance, a significant spike in volume seems to precede a rise in the Bitcoin price, suggesting that an increase in trading volume could be a leading indicator of price movement. Conversely, peaks in price do not always seem to be matched by peaks in volume, indicating that price peaks can occur without a significant change in volume.



Reply

Attach

Cancel Post Reply

-

Wei Zhao (https://northeastern.instructure.com/courses/170748/users/158006)

Feb 14, 2024

Flash Paper - Analyze Salaries of Data Science Data Description & Purpose According to the "Dat

Flash Paper - Analyze Salaries of Data Science

Data Description & Purpose

According to the "Data Science Salaries 2023" dataset on Kaggle by Arnab Chaki, we can know that this dataset is about salaries of different Data Science fields in the Data Science domain (Chaki, 2023). The data set contains data from 2020 to 2023 in different countries and regions, different job titles, and different experience levels. The link for this dataset: **https://kaggle.com/datasets/arnabchaki/data-science-salaries-2023/data (https://kaggle.com/datasets/arnabchaki/data-science-salaries-2023/data)** .

Reason for Choosing Dataset & Problems to Explore

First of all, I am looking for a job, so I chose to analyze this dataset to see what the salary is in the Data Science field. I can also know about the average salary in Data Science field, which will help me better plan my future development direction. Secondly, I think the data in this dataset is relatively new, and it can be more meaningful to analyze it. Through analyzing this dataset, I want to know which type of job has higher salary, the salary trends of different types of jobs in recent years and what kind of job is better to choose in different experience level.

Data Structure & Prepreocessing

Data Structure:

There are 11 columns and 3755 rows in this dataset. Each column are:

1. work_year: (Numeric - Double) The year in which the salary was paid.
2. experience_level: (Categorical - Character) The level of experience in the job during the respective year.
3. employment_type: (Categorical - Character) The nature of employment for the role (e.g., full-time, part-time, contract).
4. job_title: (Categorical - Character) The specific job role held during the respective year.
5. salary: (Numeric - Double) The total gross salary amount paid.
6. salary_currency: (Categorical - Character) The currency code (ISO 4217) used for the salary amount.
7. salary_in_usd: (Numeric - Double) The salary amount converted to USD.
8. employee_residence: (Categorical - Character) The primary country of residence of the employee during the respective year (ISO 3166 country code).
9. remote_ratio: (Numeric - Double) The proportion of work performed remotely, expressed as a ratio.
10. company_location: (Categorical - Character) The country where the employer's main office or contracting branch is located.

11. company_size: (Categorical - Character) The size of the company based on the median number of employees during the respective year. (Chaki, 2023)

I am interested in these variables: work_year, experience_level, job_title, salary_in_usd, remote_ratio.

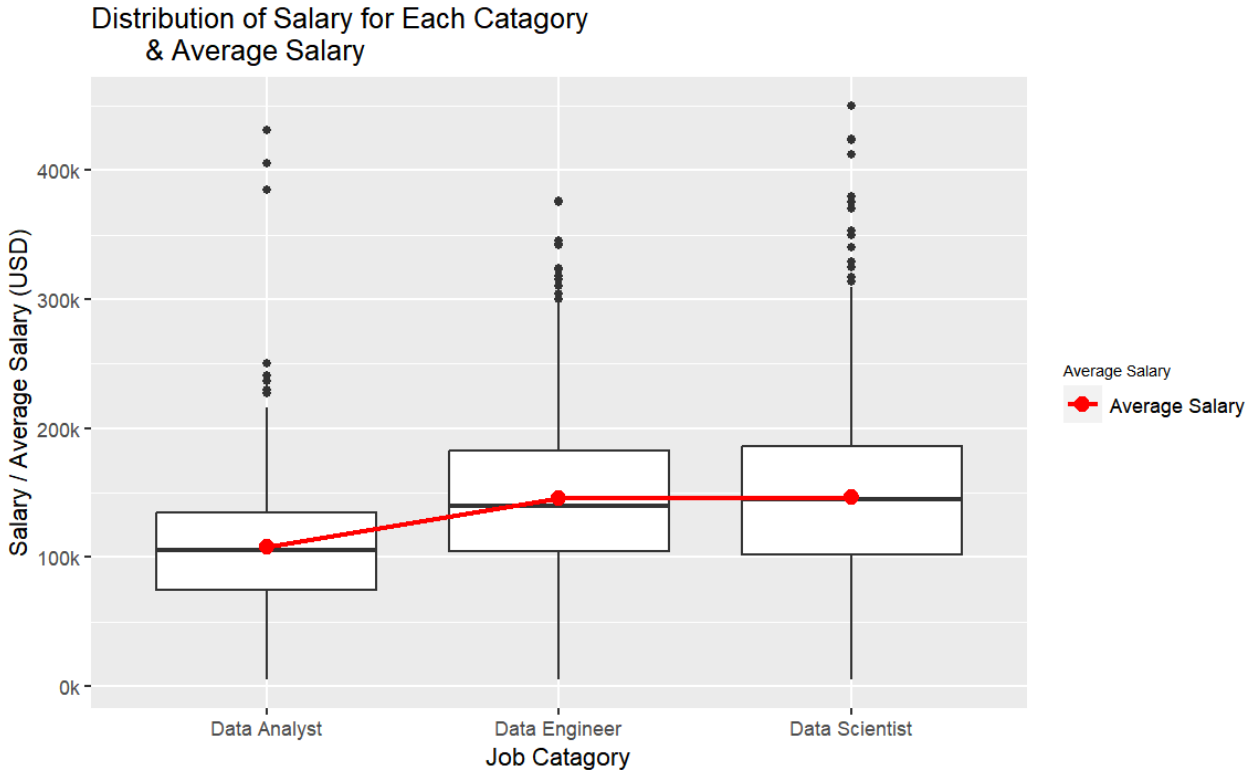
Preprocessing:

I think I need to clean and transform data for easy visualization . Following are the steps:

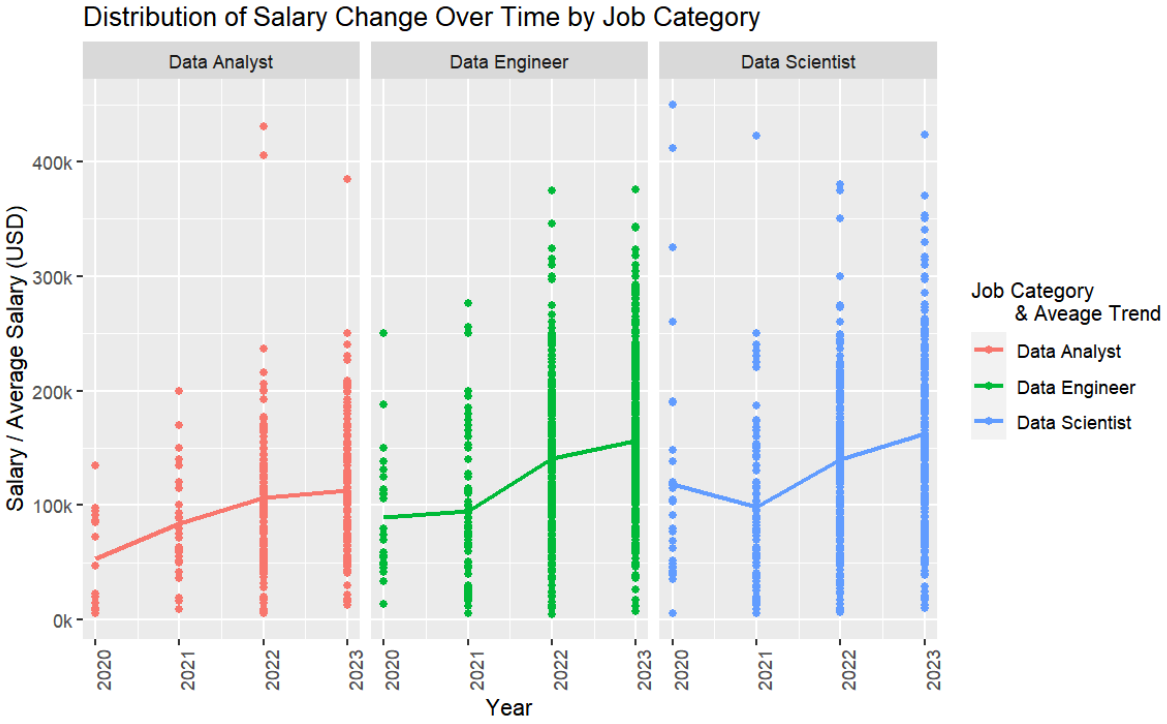
1. Due to there so many kinds of job titles and it is hard to visualize, summarize the job titles to three categories: Data Analyst, Data Scientist, Data Engineer. Mapping the job titles with job categories by create an mapping funvtion.
2. Create a mapping list for experience level without abbreviation. Replace the values in the 'experience_level' column without abbreviation.
3. Defind an order level of experience level from entry to expert. Convert experience_level to factor with specified order.
4. Filter the employee type for full time because I only want to know the distribution about full time jobs.
5. Select the columns need to use in visualization: work_year, experience_level, job_title, salary_in_usd, remote_ratio.

Visualizations & Key Findings

Following are the plots and takeaways:



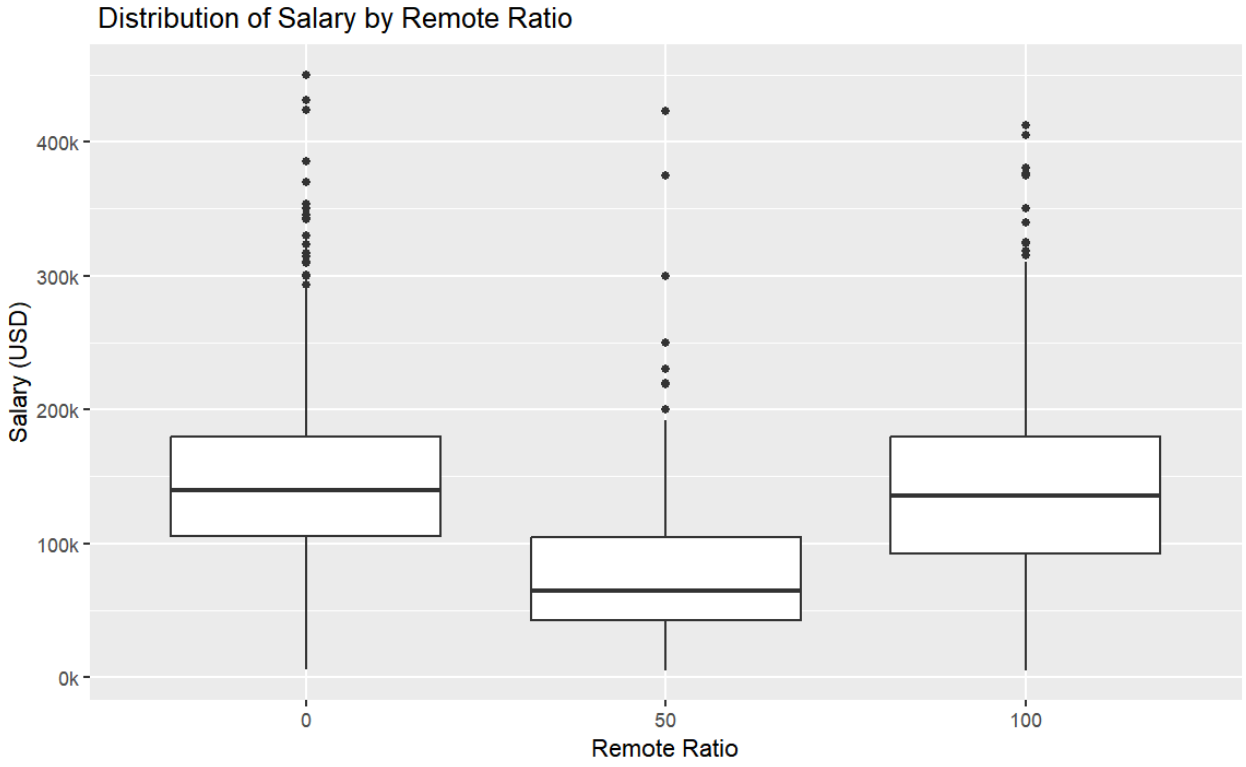
According to the plot above, I find that the job category of data scientists and data engineers have comparable salaries.The job category of data analyst have significant less salary than other two types.



According to the plot above, we can know that the salary for three types of job from 2021 to 2023, predicting the salary of all them are likely to increase in the future. The most of data analyst's annual salary are less than $250000 in 2023. Most of data scientists' and data engineers' salary are less than $300000.

According to the plot above, we can know that the average salary data analyst are less than other

two types of jobs in each level but the spread of salary is more concentrate than other two types. For expert level, it is better to find the job of data engineer. For middle level, it doesn't matter what kind of job you're looking for.



According to the plot above, we can know that the salary distribution does not change a lot comparing 0 and 100% of remote ratio. I ignore the 50% because the number of data for 50% remote ratio is more less than others, which might not be representative.

Reference:

Chaki, A. (2023). *Data Science Salaries 2023*. Kaggle. Retrieved from **https://www.kaggle.com/datasets/arnabchaki/data-science-salaries-2023/data (https://www.kaggle.com/datasets/arnabchaki/data-science-salaries-2023/data)**

Edited by **Wei Zhao (https://northeastern.instructure.com/courses/170748/users/158006)** on Feb 14 at 6:49pm

Reply

Attach

Cancel Post Reply

-

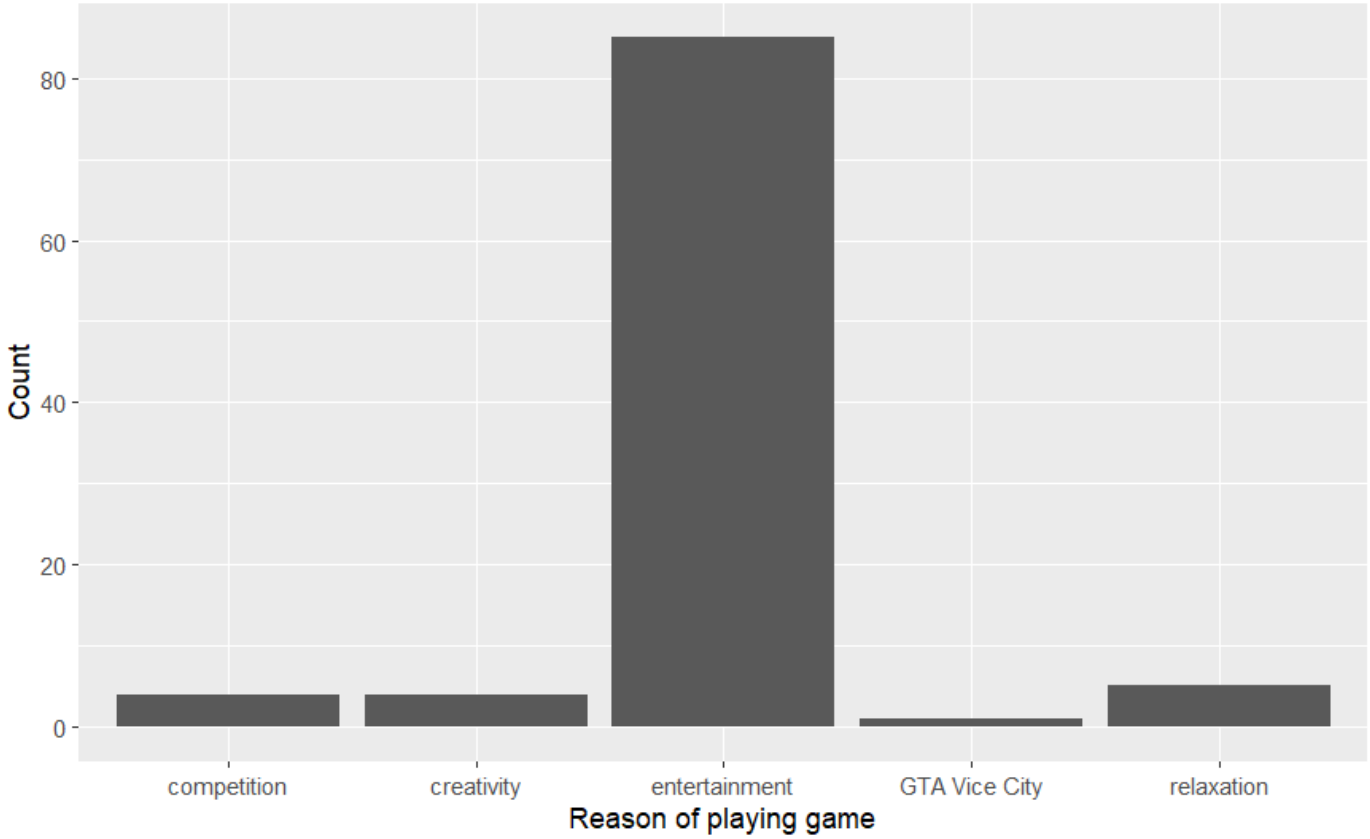
Yiduo Liu (https://northeastern.instructure.com/courses/170748/users/235612)

Feb 14, 2024

I chose to explore the Video Game Usage dataset, which I sourced from Kaggle. I was drawn to thi

I chose to explore the Video Game Usage dataset, which I sourced from Kaggle. I was drawn to this dataset because I'm also a game player and I'm curious about the reason why other people play games.

The dataset consists of variables such as age, gender, favorite games and reason of play games. The data is relatively clean, so I didn't do any preprocessing steps.



One of visualizations I created was a bar chart showing the number of people by the different reasons of playing games. This visualization shows that most people play games for entertainment, and only few people play games for other reasons like creativity and competition. Others play games

because they're maybe the super fans of a certain games.

Reply

Attach

Cancel Post Reply

-

Nikita Puri (*She/Her*) (https://northeastern.instructure.com/courses/170748/users/112995)

Feb 15, 2024

Source of Data: Centers for Disease Control and Prevention. (n.d.). Provisional covid-19 deaths by

Source of Data: Centers for Disease Control and Prevention. (n.d.). ***Provisional covid-19 deaths by sex and age***. Centers for Disease Control and Prevention. https://data.cdc.gov/NCHS/Provisional-COVID-19-Deaths-by-Sex-and-Age/9bhg-hcku/about_data

The data was updated as of September 27, 2023, and it shows the provisional Covid-19 deaths by sex and age as stated in the title. It consists of the following variables some of which are characters while others are integers:

- Dates

- State

- Sex

- Age Group

- Covid-19 Deaths

- Total Deaths

- Pneumonia Deaths

- Pneumonia and Covid-19 Deaths

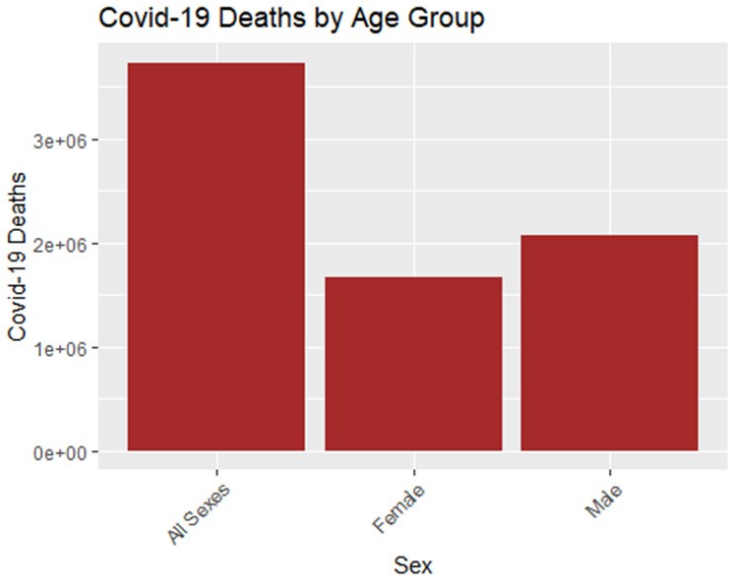
- Influenza Deaths

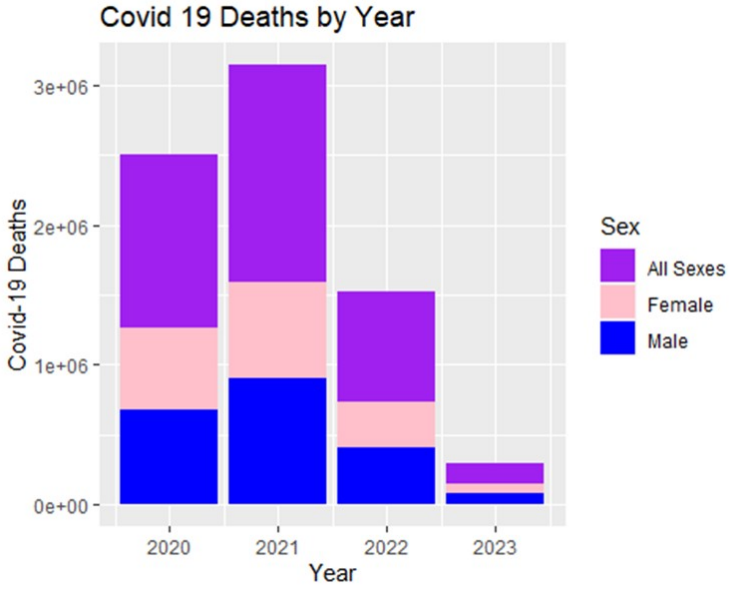
- Pneumonia, Influenza, or Covid-19 Deaths

The data is straightforward and provides information on the number of deaths that have been caused by Covid-19, Influenza, and Pneumonia or a combination of these. It looks at the numbers in terms of sex and age group. I chose this data set because Covid-19 has been prevalent since 2020. It was a global pandemic which forced the entire world to go on lockdown impacting economies and in turn, exacerbating the circumstances of middle class and poor families. I had not seen a data which emphasized the number of deaths by sex and age prior to this assignment. Therefore, I found it compelling to explore this data set and visualize the statistics behind the conversations people often have about Covid and its impact on various people. However, daily conversations are different from the real numbers which are representative of the reality around us.

Furthermore, not long before starting this assignment, I was having a conversation with one of my clients from work. They had a health sciences background and spoke about the prevalence of more people becoming sick as of recently. They emphasized that while the Covid-19 vaccines have been effective for a few years now, virus strains evolve, and people are still succumbing to the symptoms that arise. However, we both agreed that it is not talked about anymore, at least not as much as in 2020 or 2021. With each passing year, less and less people wear masks which make us all vulnerable to the fast spread of viruses. The lack of media coverage on the dangers of evolving viruses can lead people to believe that they do not exist or may not have serious implications. This was one of the reasons I wanted to explore this data as it pertains to Covid-19. I wanted to take a closer look at whether Covid-19 impacts people in the same level of severity as it did prior to the release of the vaccines. In other words, are Covid-19 deaths declining with the passing of each year since the vaccines?

The data set is large and consists of 16 columns, each representing different variables of interest as listed above. There are a total of 137,700 entries but despite its size, the data was fairly easy and quick to load into R. As part of preprocessing, I created the "CountNA" function which counts missing values by row or column. This was a function that we created for our first homework assignment. I was able to utilize this very helpful function here in order to count the number of missing values since there were a large number of them. Then, I omitted all NAs using the pre-existing "na.omit" function to ensure calculations and visualizations did not contain any discrepancies. Because the data had been tidied ahead of being loaded into R on the source website, there were no additional steps that needed to be taken. The data was also already separated based on character or numeric features, which made pre-processing very simple.





The first graph is looking at the number of deaths by sex. Clearly, there are far more males who have died due to Covid-19 compared to females. The second graph shows a similar breakdown of Covid-19 deaths but has an additional component of year. This gives more insight into how many deaths took place in each given year while also showing the breakdown of the sexes which is coded in different colors. The second graph shows that there were far more Covid-19 deaths in 2021 than in the years 2020, 2022, or 2023. In all of the years, more males have died than females. Prior to looking at this data, I had not really thought about the distinctions in the number of deaths between men and women. However, this really gives insight into the fact that more men have died due to Covid-19 than women. This can lead to more insightful questions encouraging research on the factors that may cause more deaths in men than women, which is outside of the scope of this particular data set.

Additionally, the question that I had initially sought to answer regarding whether deaths due to Covid-19 were declining with the passing of each year, is also answered by this visualization. Clearly, Covid-19 deaths have declined significantly since the introduction of the pandemic in 2020. Covid-19

deaths were extremely high in 2020 and 2021, reaching their peak in the year 2021. This may be attributed to various factors including the emergence of new variants of the virus, vaccination rates, and global response at the time. These visualizations lead to more curiosity within the topic and encourage further research into these factors. Moreover, as mentioned above, there can be further investigation on the impact of variants and the evolving nature of viruses. The next steps could look into impact of Covid-19 variants and their prevalence as of recent years.

Reply

Attach

Cancel Post Reply

-

Aditya Sairam Govindan (https://northeastern.instructure.com/courses/170748/users/306162)

Feb 15, 2024

Flash Paper - Wine Reviews. Overview of the Dataset : For the second assignment, I selected the

Flash Paper - Wine Reviews.

Overview of the Dataset :

For the second assignment, I selected the **"Wine Reviews"** dataset, a publicly available dataset accessible on Kaggle. The Wine Reviews dataset offers a comprehensive repository of information encompassing diverse facets of wine production globally. This extensive dataset provides valuable insights into individuals' preferences, reflecting their distinct taste palettes and cultural inclinations towards various wine types. By delving into the details of wine varieties, pricing structures, appellations, and more, this dataset becomes a rich source for understanding the intricacies of people's preferences for alcoholic beverages. Analyzing this dataset unveils a profound opportunity to decipher consumer behavior and discern the factors influencing their choices across different regions and demographics.

Description about each column:

Wine: Name of the wine.

Winery: Name of the winery.

Category: Type of wine.

Designation: designation of wine.

varietal: type of grape.

appellation: region of wine.

alcohol: alcohol content.

price: price in dollars.

rating: rating from reviews.

reviewer: name of reviewer.

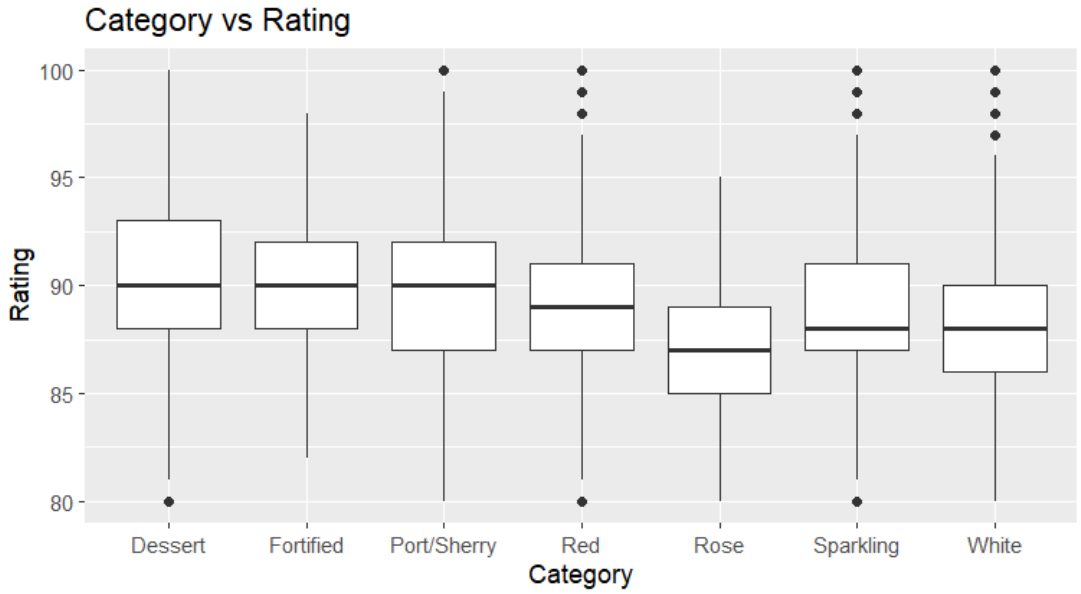
Pre-processing and Text Cleaning:

Upon meticulous examination of the dataset, several irregularities were identified. Notably, the "**appellation**" column contained values that were incongruous, specifically '***Buy Now***' and '***Drizly-Vivino***'. These values were removed. In the "**price**" column, the presence of the '**$**' symbol warranted pre-processing. The column's data type was adjusted to **numeric**, with the '**$**' symbol appropriately removed. Similar data type evaluations and adjustments were performed for other columns.

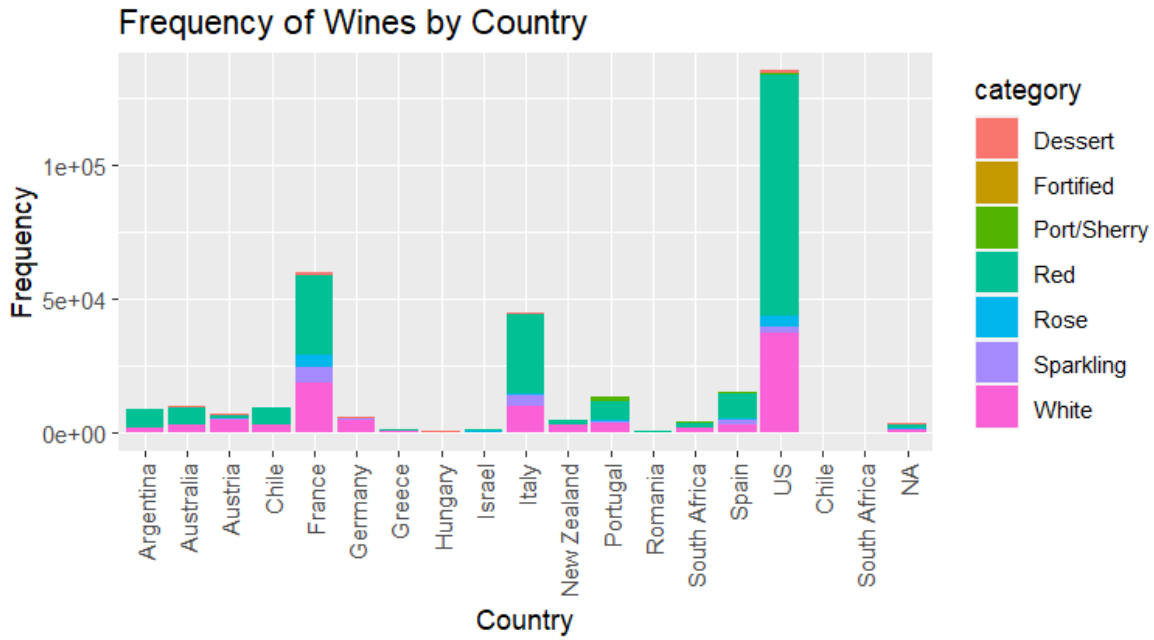
Additionally, the dataset exhibited a considerable number of '**NaN**' values. To address this, a custom imputation function, **imputeNA** (implemented in Home work-1) , was employed to enhance accuracy and suitability for subsequent analysis.

Analysis and Plots:

An initial exploratory analysis focused on understanding the distribution of reviews for each wine type. This plot revealed nuanced preferences among consumers. Notably, wine types such as '**Dessert**', '**Fortified**', and '**Port/Sherry**' exhibited closely aligned rating distributions, with mean scores hovering around **90**. In contrast, '**Rose**' emerged with the lowest mean rating at **87.5**.



A subsequent exploration delved into wine production, dissected by countries. The plotted data underscored the dominance of the **USA** in wine production, surpassing other nations significantly, followed by **France**. Noteworthy was the prevalence of '**Port/Sherry**' as the most commonly produced wine type across diverse countries.



This refined dataset, coupled with insightful visualizations, lays the groundwork for a sophisticated analysis of global wine preferences and production trends.

Citation :

Dataset Description : The dataset is created by scrapping wine reviews web pages. There are over 320k wines and their associated features.

License : Data files © Original Authors.

Link to Dataset : https://www.kaggle.com/datasets/samuelmcguire/wine-reviews-data/data (https://www.kaggle.com/datasets/samuelmcguire/wine-reviews-data/data) . (https://www.kaggle.com/datasets/samuelmcguire/wine-reviews-data/data.)

Edited by Aditya Sairam Govindan (https://northeastern.instructure.com/courses/170748/users/306162) on Feb 15 at 10:25am

Reply

Attach

Cancel Post Reply

-

Karthikeyan Sugavanan (https://northeastern.instructure.com/courses/170748/users/306200)

Feb 15, 2024

Q1. Describe the dataset and where it comes from (making sure to cite the data source). Explain w

Q1. Describe the dataset and where it comes from (making sure to cite the data source). Explain why you chose this dataset and what questions you wanted to explore in your visualization.

A1.

"The Tax Burden on Tobacco, 1970-2019," a federal dataset from the Centers for Disease Control and Prevention (CDC) website, was used for this analysis. This dataset offers thorough details on the taxes imposed on tobacco products spanning almost fifty years. The information provides information

on state and federal tax burdens as well as the amount of money the government generates from the sale of tobacco products.

This dataset was chosen due to it is important to comprehending the effects of tobacco use on the economy and public health. By exploring the tax burden on tobacco, one can examine the financial incentives for reducing tobacco usage and assess the effectiveness of policies aimed at curbing smoking rates. Additionally, this dataset allows for insights into regional disparities in tobacco consumption and expenditure, highlighting potential areas for targeted interventions.

In visualizing this data, the aim is to depict the trends in tobacco taxation over time and across different geographical regions. Specifically, I seek to visualize the variation in tax revenue collected by each state, the proportion of federal versus state taxes, and the overall gross revenue generated. Furthermore, the visualization will shed light on patterns of tobacco expenditure, enabling the identification of states with higher rates of tobacco consumption and potential correlations with socioeconomic factors.

Dataset : **https://data.cdc.gov/Policy/The-Tax-Burden-on-Tobacco-1970-2019/7nwe-3aj9/about_data (https://data.cdc.gov/Policy/The-Tax-Burden-on-Tobacco-1970-2019/7nwe-3aj9/about_data)**

Q2. Describe the structure of the dataset and the variables of interest. Describe any preprocessing needs (tidying, cleaning, transformation, etc.) and describe the steps you took to perform the preprocessing.

A2.

The dataset comprises 17 columns and 15,300 rows, with key variables of interest including "LocationDesc," "Year," "SubMeasureDesc," "Data_Value," "Data_Value_Unit," "Data_Value_Type," and "GeoLocation." These variables provide insights into the location, year, specific measures related to tobacco taxation, and corresponding data values.

Preprocessing procedures were carried out in order to improve the dataset's utility for exploratory analysis. The data in the column "SubMeasureDesc"—such as Average Cost per pack, Cigarette Consumption (Pack Sales Per Capita), Federal and State tax as a Percentage of Retail Price, Federal and State tax per pack, Gross Cigarette Tax Revenue, and State tax per pack—would be better utilized if they were presented as separate columns.

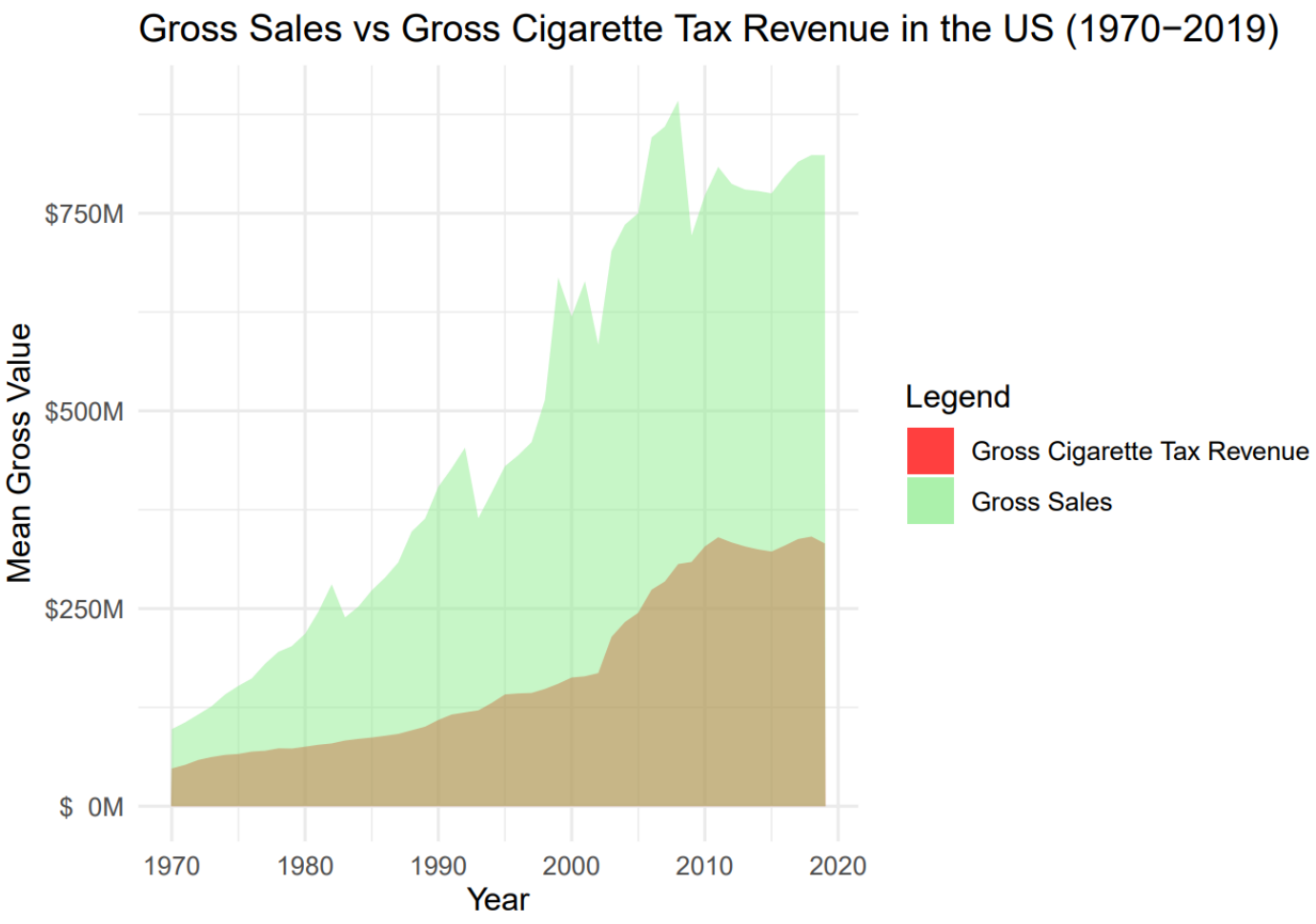
The "pivot_wider()" function from the "tidyverse" package was applied to the "SubMeasureDesc" column in order to do this. Furthermore, a new column called "Gross sales" was included. It was determined by utilizing the data from "Gross Cigarette Tax Revenue" and "Federal and State tax as a percentage of Retail Price."

Beyond these preprocessing steps, the data was found to be already tidy and clean, necessitating minimal further tidying or cleaning. This allowed for streamlined exploratory analysis of the dataset, facilitating insights into trends and patterns related to tobacco taxation across different locations and years.

Q3. Present at least 1 figure that is interesting to you and describe your observations and any key takeaways from the visualization and your exploration of the dataset.

A3.

(a)



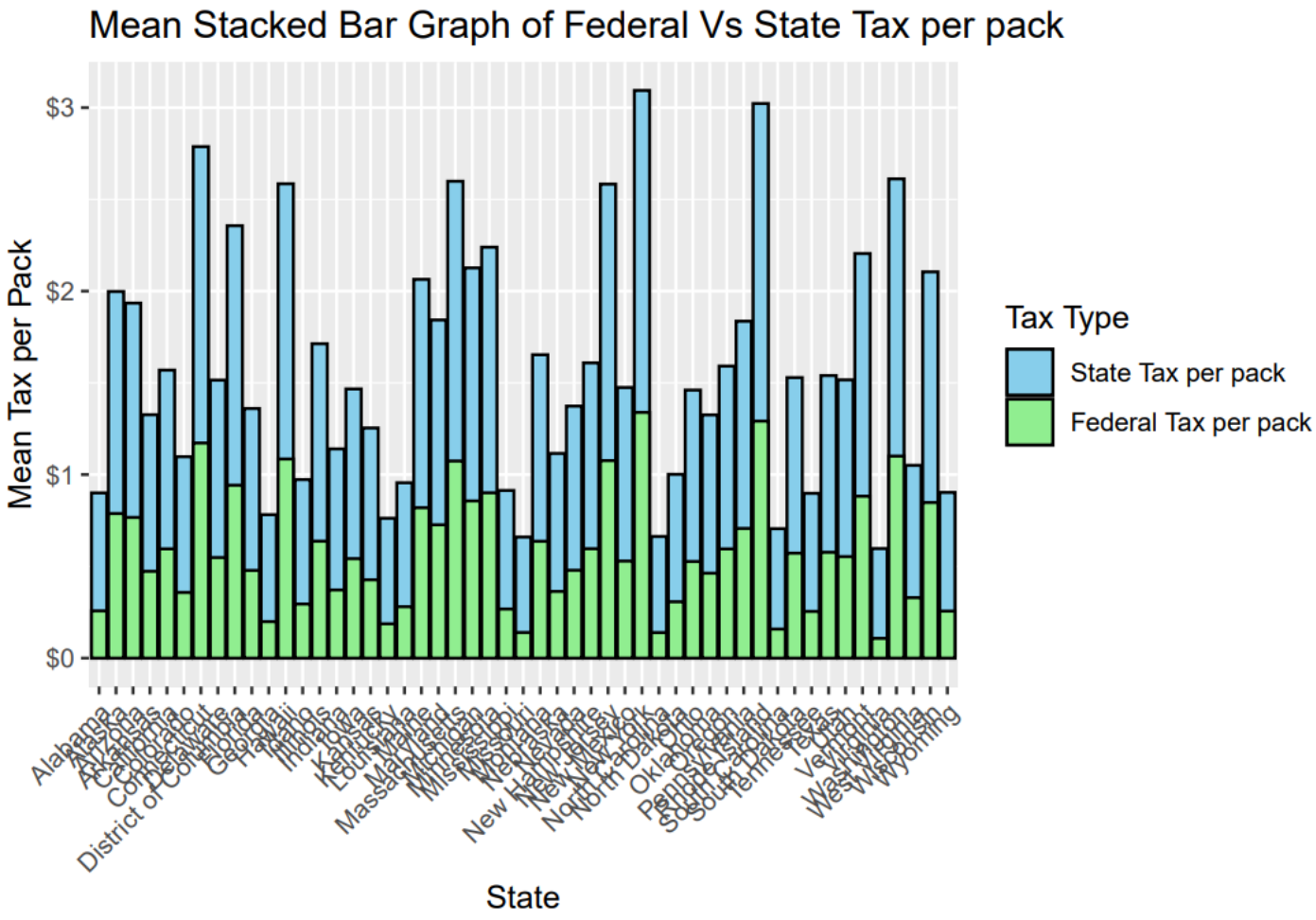
The graph depicts the Gross Sales of cigarette in the US in the time period 1970-2019 , and also the the Gross tax revenue collected by the Federal and state together.

Some noticeable infrences from the graph include, Gross cigarette sales have experienced a notable decline since 1991, dropping from around $600 million to approximately $200 million by 2019. This downward trend can be attributed to various factors such as heightened awareness regarding the health hazards of smoking, escalating cigarette taxes, and the emergence of alternative tobacco

products like e-cigarettes. Concurrently, gross cigarette tax revenue has also decreased, albeit at a slower pace compared to sales. Despite declining sales, tax revenue fell from around $200 million in 1991 to about $100 million in 2019, primarily due to escalating cigarette taxes and potential issues like smuggling and tax evasion.

Furthermore, the dataset illustrates a consistent increase in the average price of cigarettes since 1970, influenced by factors like inflation, escalating taxes, and legal costs associated with litigation. This upward trajectory in prices likely contributed to a reduction in cigarette consumption. Altogether, these trends reflect a diminishing popularity of cigarette smoking in the United States, driven by heightened awareness of health risks, increased taxation, and the advent of alternative tobacco products, collectively shaping the landscape of tobacco consumption habits.

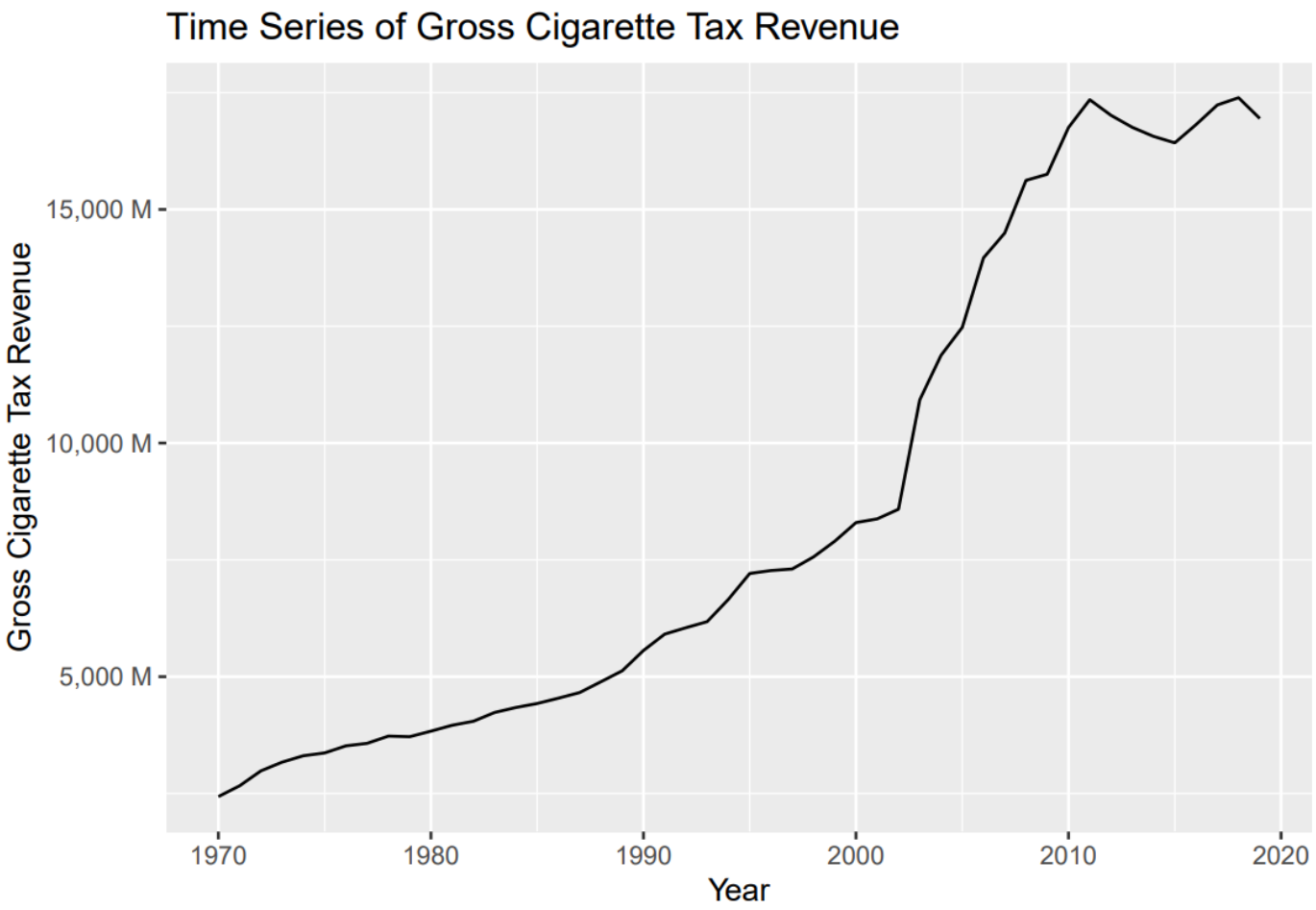
(b)



The graph illustrates the mean Tax Revenue collected separately by State and Federal entities for the sale of each cigarette packet. Notably, the Federal tax per pack consistently surpasses the state tax per pack across all listed states, with a minimum federal tax of $1 per pack while state taxes range from zero to over $2 per pack. The considerable variance in state tax per pack is evident, with states like Alaska and Massachusetts imposing taxes exceeding $2 per pack, contrasting with states like North Carolina and Virginia, which levy no state tax per pack.

Examining the total tax burden per pack reveals significant disparities among states. Massachusetts and Alaska exhibit the highest total tax per pack, stemming from the combination of federal and state taxes. Conversely, North Carolina and Virginia record the lowest total tax per pack due to the absence of state taxes, emphasizing the influence of both federal and state taxation policies on the overall taxation landscape.

(c)



The graph illustrates the consistent growth in gross cigarette tax revenue in the United States from 1970 to 2019. Over this period, revenue surged from approximately $2 billion to nearly $15 billion, indicating a substantial sevenfold increase, even when adjusted for inflation. However, the rate of growth has tapered off in recent years, particularly around the year 2000. Various factors could contribute to this slowdown, including declining smoking rates, heightened cigarette smuggling activities, or alterations in tax policies.

Furthermore, the graph hints at a potential cyclical pattern in the growth of cigarette tax revenue, with fluctuations occurring approximately every four years. This cyclicality may be influenced by economic shifts, alterations in tax rates, or other external factors. Understanding these patterns and their underlying drivers is crucial for policymakers and stakeholders in the tobacco industry to effectively anticipate and respond to changes in cigarette tax revenue over time.

Reply

Attach

Cancel Post Reply

Wren Warren (*They/Them*) (https://northeastern.instructure.com/courses/170748/users/68901)

Feb 15, 2024

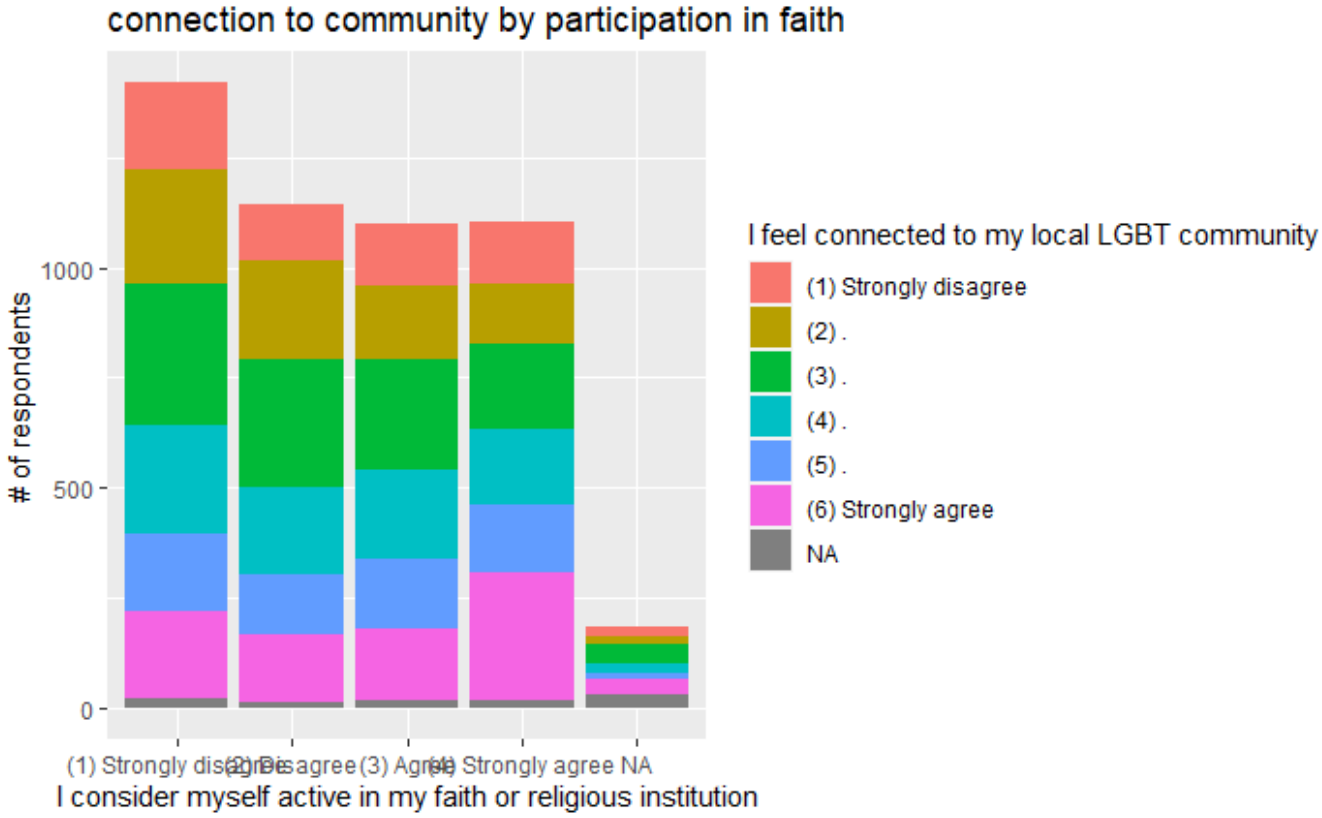
The dataset I chose to work with for homework 2 is from the Social Justice Sexuality project, gathe

The dataset I chose to work with for homework 2 is from the Social Justice Sexuality project, gathered from the Resource Center for Minority Data. The dataset consists of interview answers from over 5000 participants given on the topics of race, sexuality, community, mental health, and faith by Black and Indigenous People of Color (BIPOC) members of the Lesbian, Gay, Bisexual, Transgender, Queer, Intersex and Asexual (LGBTQIA) community. Respondents were from across the US.

I chose this dataset because I believe that intersectionality is important and it is necessary to examine the perspectives of those minorities who have intersecting disadvantages from both ethnic/racial background and gender and sexual minority status. Such perspectives can give us helpful insights into the ways in which our communities can fail their members, and thus can provide potential opportunities to improve our communities for the good of all (including the most disadvantaged members of those communities). I wanted to explore how the intersection of queerness and racial minority status affected participants understandings of their communities with regard to familial connection, religion, and overall mental health.

The preprocessing done on this data was extensive, most notably transforming race and gender into single variables as opposed to a separate variable for each gender and race option, and changing variable names to include information identifying that variable such with key features (as opposed to the variable being named “Q16B” for example). I also made sure variables were encoded as factors and removed variables I considered extraneous or which had been blanked for anonymity.

The attached figure demonstrates an interesting relationship between activity in faith-based or religious practices and a feeling of connection to the participant's local LGBT community, where there is a slight trend of people who more strongly agree they are active in their faith being more connected to their LGBT community. This flies against the assumptions of some that LGBT community would act as a replacement for a faith-based community, and indicates that those more active in one facet of their lives may be more connected to others in other facets of their lives, at least in racial and gender/sexual minority groups in the United States.



Reply

Attach

Cancel Post Reply

-

Arpita Gupta (https://northeastern.instructure.com/courses/170748/users/265169)

Feb 15, 2024

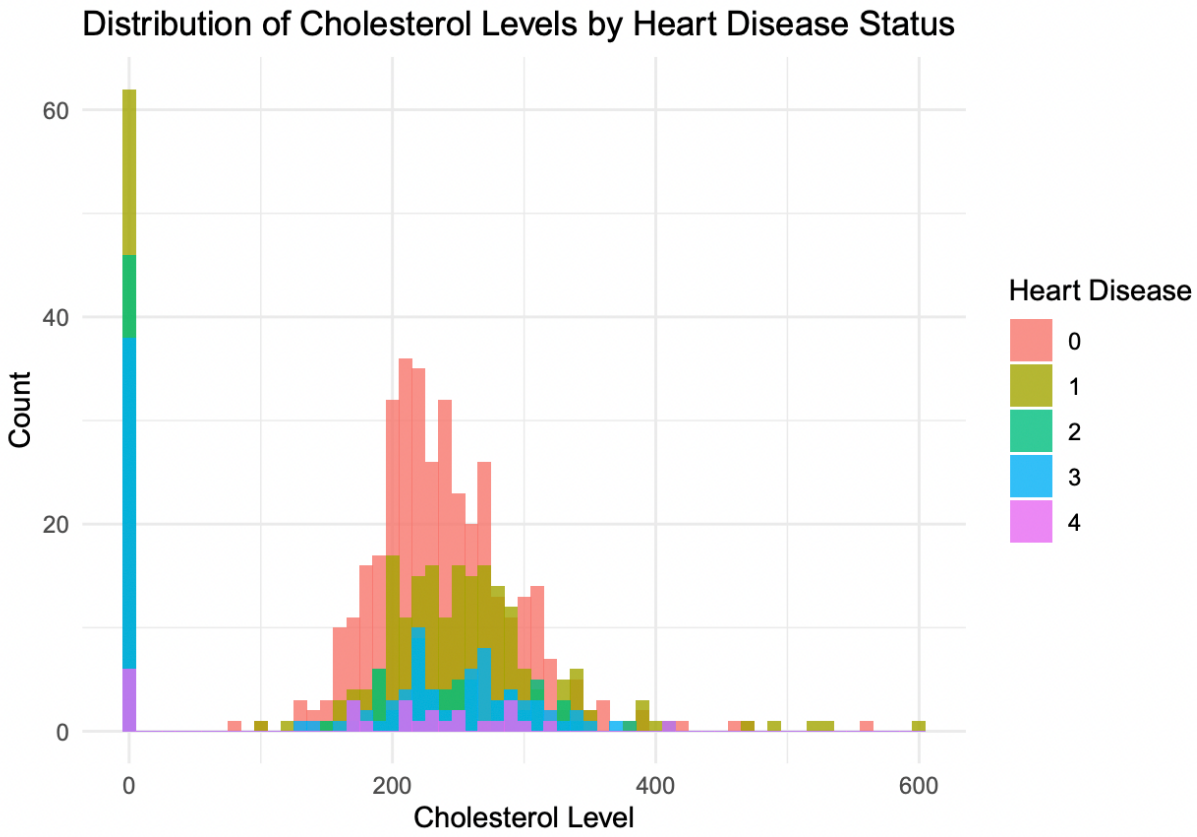
Dataset Source: Dataset 1. Dataset Description: The dataset selected originates from Kaggle and e

Dataset Source: **Dataset (https://www.kaggle.com/datasets/redwankarimsony/heart-disease-data)**

1. **Dataset Description:** The dataset selected originates from Kaggle and encompasses data concerning various factors linked to the presence or absence of heart disease in patients. It encompasses variables such as age, sex, chest pain type, resting blood pressure, serum cholesterol level, and fasting blood sugar level. The primary target variable denotes the presence or absence of heart disease. This dataset was chosen due to its pertinence in public health and its capacity to unveil insights into the factors correlated with heart disease.

2. **Preprocessing And Cleaning:** In terms of structure, the dataset entails both categorical and continuous variables. Before commencing any analysis, preprocessing steps were imperative to ensure data integrity. This encompassed scrutinizing for missing values, addressing outliers, and potentially encoding categorical variables. Standard data preprocessing techniques, including imputation for missing values, outlier identification and treatment, and categorical variable encoding, were employed as necessary.

3. **Key Visualization:** Among the visualizations generated, a notable one is a histogram depicting the distribution of cholesterol levels categorized by heart disease status. The histogram illuminates discernible patterns in cholesterol levels among individuals with and without heart disease. Intriguingly, there's a noticeable prevalence of elevated cholesterol levels among individuals with heart disease compared to those without. This observation underscores the established correlation between high cholesterol levels and cardiovascular health. Further analysis could delve into potential correlations between cholesterol levels and other risk factors in the dataset, offering valuable insights for preventive measures against heart disease.



Overall, this exploratory analysis sets the stage for deeper investigation into the complex interplay of factors contributing to heart disease, potentially informing strategies for prevention and treatment.

Edited by Arpita Gupta (https://northeastern.instructure.com/courses/170748/users/265169) on Feb 15 at 10:19am

Reply

Attach

Cancel Post Reply

-

Manjusha Motamarry (https://northeastern.instructure.com/courses/170748/users/

306153)

Feb 15, 2024

Name: M.G.Manjusha , Section 3 1.The data set I have chosen is taken from the following source:

Name: M.G.Manjusha , Section 3

1.The data set I have chosen is taken from the following source:

https://hub.tumidata.org/dataset/airpollutionbeforeandafterlockdownindelhi_delhi (https://hub.tumidata.org/dataset/airpollutionbeforeandafterlockdownindelhi_delhi)

The dataset includes data about pollution before and after lockdown in India. It includes pollutant data recorded at different locations over time. The data is recorded every one hour from 12 AM to 11PM. This dataset was selected due to its relevance to current environmental concerns and its potential for insights into the impact of lockdown measures on air quality.

Here are the questions I aimed to explore through visualization:

- Effect of Lockdown on Pollutant Levels: I wanted to investigate how the implementation of lockdown phases affected pollutant levels. By visualizing pollutant values across different lockdown phases, I sought to observe any discernible trends or changes in pollution levels before, during, and after lockdown periods.
- Spatial Distribution of Pollutants: Through visualization, I aimed to analyze the spatial distribution of pollutants across different states. This involved exploring variations in pollution levels among states and identifying regions with consistently high or low pollutant levels.
- Temporal Trends in Pollution: I wanted to examine temporal trends in pollution levels over the duration of the dataset. By visualizing pollutant values over time, I aimed to identify any seasonal patterns or long-term trends in pollution levels.
- Comparative Analysis: Additionally, I sought to conduct a comparative analysis of pollutant levels before and after lockdown. This involved comparing pollutant values between different phases of lockdown to assess the effectiveness of lockdown measures in mitigating pollution.

2. Structure of the dataset:

The dataset includes the following columns:

- datetime: Gives the date and the time of when the value was recorded.Each day has recordings from 12 AM to 11 PM
- id : Gives the unique id for each of the cities and states.
- name : Name of the sector
- longitude

- latitude
- live: A boolean of True or False. If the data was recorded live it has a True value else it has False value.
- cityid : name of the city
- stateid : name of the state
- PM2.5 : Pollutant:Particulate Matter with a diameter of 2.5 micrometers or less, which can penetrate deep into the respiratory system.
- PM10:Pollutant: Particulate Matter with a diameter of 10 micrometers or less, which can be inhaled into the respiratory system and cause health issues.
- NO2:Pollutant: Nitrogen Dioxide, a harmful gas primarily emitted from vehicle exhaust and combustion processes, contributing to air pollution and respiratory problems.
- NH3:Pollutant:Ammonia, a compound released from agricultural activities, livestock waste, and industrial processes, contributing to air pollution and environmental concerns.
- SO2:Pollutant:Sulfur Dioxide, a gas released from burning fossil fuels containing sulfur, contributing to air pollution and acid rain formation.
- CO: Pollutant:Carbon Monoxide, which can cause harmful health effects by reducing oxygen transport in the bloodstream.
- OZONE: Pollutant: Ground-level Ozone, causing respiratory issues and environmental damage.

Pre-Processing steps taken:

Tidying the data:

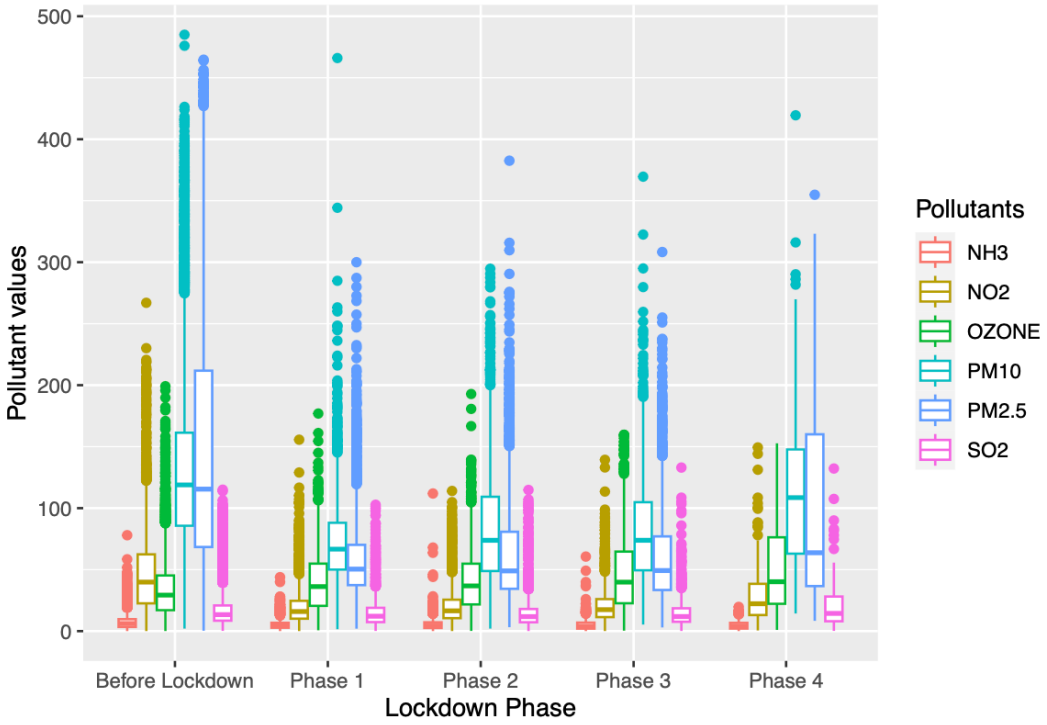
- The dataset has 15 columns, out of which 4 columns can be discarded. The columns "live", "name", ”longitude”, and “latitude” are discarded as we do not use them.
- The existing datetime column is changed to represent only dates and not times.
- The data recorded from 12 AM to 11 PM takes up 11 rows. I converted it to one column by taking the mean of the pollutant values of all 11 hours.
- NA values are omitted.
- New column names “Lockdown Phase” is added to the dataset which tells which phase of the lockdown was going on according to the date given.

The lockdown in India was phase wise as below:

- 1st Jan 2020 - 24th March 2020 - Before lockdown
- 25th March 2020 - 14th April 2020 - Lockdown Phase 1
- 15th April 2020 - 3rd May 2020 - Lockdown Phase 2
- 4th May 2020 - 17th May 2020 - Lockdown Phase 3

- 18th May 2020 - 31st May 2020 - Lockdown Phase 4

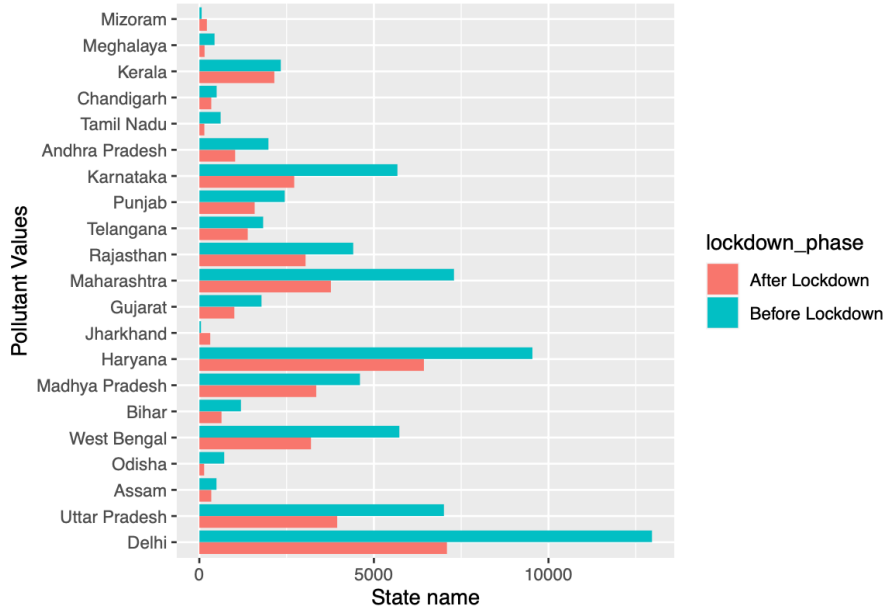
3. Observations and key takeaways



In the above plot, we can see the pollutant's minimum, maximum, and median values. All the pollutants are color coded.

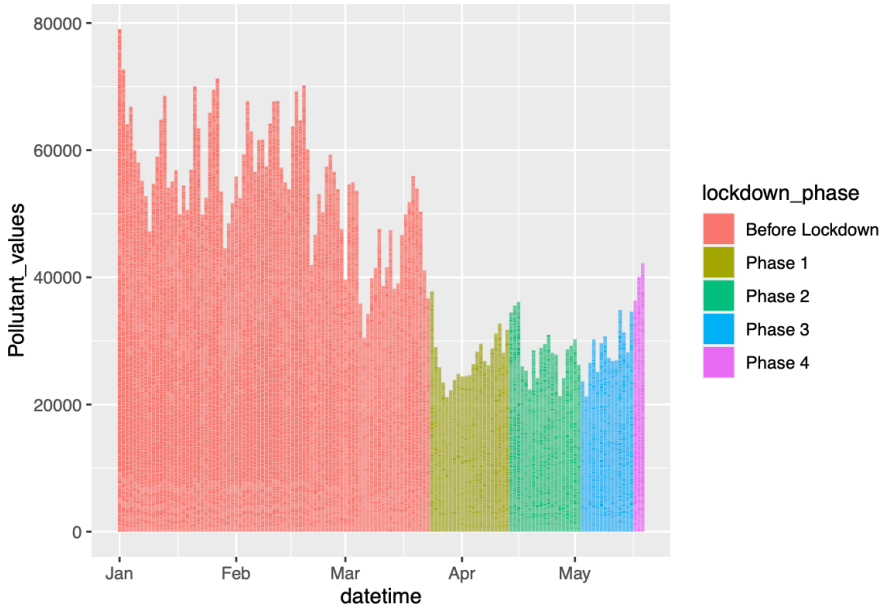
Observations:

- NH3 pollutant is less compared to other pollutants before and after lockdown(Phase 1,2,3,and 4)
- The pollutants which are a lot are: PM10 and PM2.5
- All pollutants except NH3 and SO2 had significant reduction after lockdown compared to before lockdown.



From the above plot, notable points are:

- Almost every state's pollution has reduced after the lockdown.
- Jharkhand and Mizoram are the only states that are not inclining towards the trend of decrease in pollution after lockdown.
- Delhi is the most polluted state before and after lockdown.
- Harayana is the second most polluted state before and after lockdown.
- Jharkhand is the least polluted state before lockdown.
- Odisha is the least polluted state after lockdown.



- As we can see from the above plot, the pollution decreased drastically after lockdown. In this plot we can see the difference as each month goes by.
- A point to be noted is that, At the end of each phase, we see an increase of pollutants compared

to the least value in that phase of lockdown.
- We can infer that the trend of pollutant values is not continuously decreasing.
- At the beginning of each phase of lockdown including before lockdown, we see a drastic fall in the pollutant values. However, this trend does not hold true for Phase 4.

Reply

Attach
Cancel Post Reply

Sai Nikhil Kunapareddy (https://northeastern.instructure.com/courses/170748/users/309623)

Feb 15, 2024

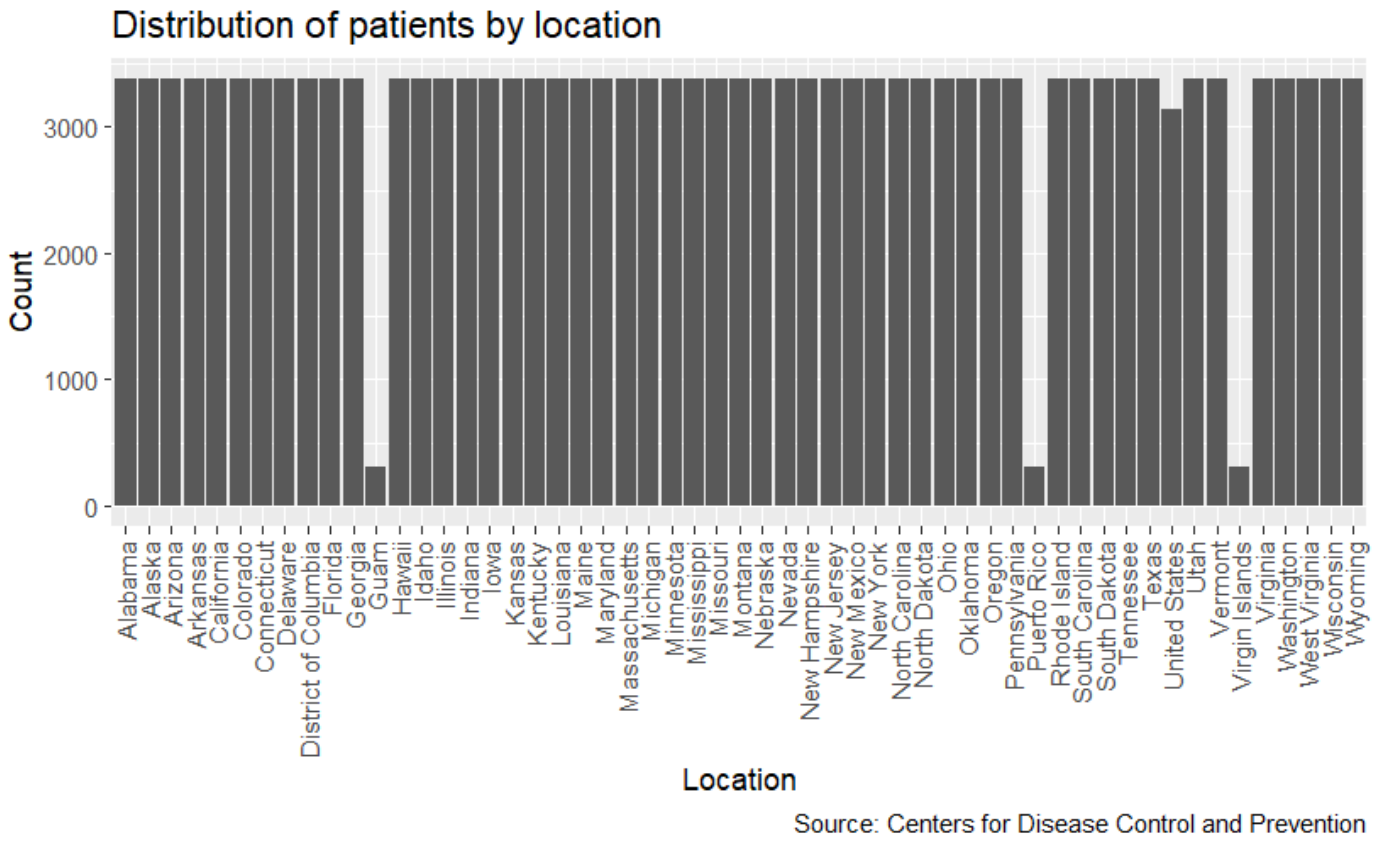
The analyzed data set is named ‘US Chronic Disease Indicators: Cancer’ taken from the ‘Centers c

The analyzed data set is named ‘US Chronic Disease Indicators: Cancer’ taken from the ‘Centers of Disease Control and Prevention’ website. The following is a link to access the same: link (https://data.cdc.gov/Chronic-Disease-Indicators/U-S-Chronic-Disease-Indicators-Cancer/u9ek-bct3/about_data) . I chose this dataset because cancer has been one of the leading causes of death for decades now. Analyzing this data might give me some understanding about the influence of patient demographics on the inception of this disease. I am curious to look into any skewness in patient data with respect to a patient location or age that is prevalent in relation to being diagnosed with cancer.

The dataset in discussion contains 33 columns and 176,339 observations. Attributes present in the dataset are: dates of diagnosis of the disease and mortality of the patient, geographical location information, type of cancer, few stratification categories that divide patients based on certain demographics. Unfortunately not all stratification columns are complete, hence only the information related to the gender and ethnicity of the patient is considered for this analysis.

The preprocessing steps taken to clean the dataset for analysis are as follows. glimpse(), summary(), head() and tail() functions are used to understand the range and categories of data present in the data set. Number of 'NA' values of each column are counted to understand the completeness of the available data. As already mentioned, few stratification category columns are completely empty, hence the columns with 'NA' values throughout all observations are removed.

I have analyzed the distribution of patients across different locations within the US, the ethnicity they come from and the gender they identify as. Below is a snapshot of one of the results.



From the above graph, the number of the patients seem to be more or less equally distributed in almost all the locations with few exceptions. This is an unexpected outcome and this trend is followed in other distributions with respect to race and gender as well. I wouldn't derive any concrete reasoning from the above data set for the same reason. Instead I would be interested in understanding the data collection and handling methods used to manipulate this data, to make sure that the dataset is reliable.

Reply

Attach

Cancel Post Reply

-

Chengbin Huang (https://northeastern.instructure.com/courses/170748/users/262574)

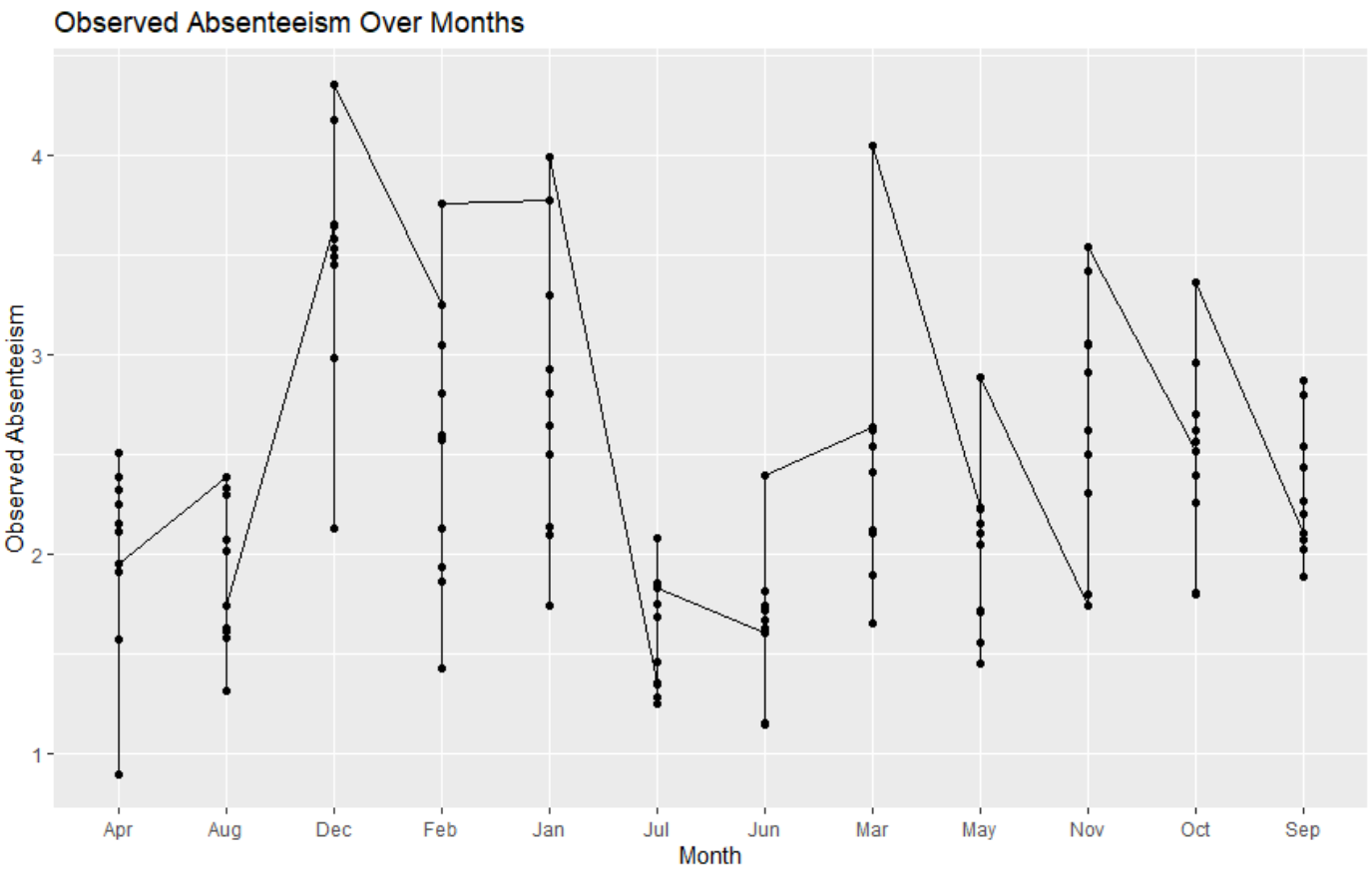
Feb 15, 2024

Dataset description and source: This dataset contains statistics on workplace absences related to a

Dataset description and source: This dataset contains statistics on workplace absences related to absences caused by different health and medical issues. The source of the data is the National Institute for Occupational Safety and Health (NIOSH), which monitors absences reported by full-time workers due to illness, injury or other medical problems, known as health-related workplace absences. I chose this dataset because I am interested in workplace health and occupational safety and wanted to analyze absence data to understand health conditions and possible issues in the workplace.

Dataset structure and variables: This dataset contains the following variables: HHS Region (regional division of the U.S. Department of Health and Human Services), Month (month), Observed (observed absences), Observed LCL (observed absences) lower limit), Observed UCL (upper limit of observed absences), Expected (expected absences), Expected LCL (lower limit of expected absences), Epidemic Threshold (epidemic threshold), Alt Text (alternative text). I will focus on the absence rate. During data preprocessing, missing value processing, data type conversion, and data rearrangement are required to ensure the accuracy and consistency of the data.

Visual analysis: I will use the ggplot2 package to create a line chart to visualize the observed absences for each month in each region. This will help me get a more visual understanding of absences in different regions in different months, and possibly spot anomalies in certain months or regions. By looking at the discount chart plot, I can identify areas or months with higher absenteeism rates, providing clues for further analysis and research.



Chengbin Huang (https://northeastern.instructure.com/courses/170748/users/262574)

Feb 15, 2024

By observing the absenteeism rate each month, we can find the following: Seasonal variati

By observing the absenteeism rate each month, we can find the following:

Seasonal variations: It may be observed that absenteeism rates are significantly higher in some months than in others, which may be related to seasonal factors. For example, you may experience higher absenteeism rates in the winter because this is the season for influenza and other respiratory illnesses. Holiday effect: Absence rates in certain months may change due to holidays. For example, during the holidays, employees may be more likely to take time off or take time off, leading to increased absenteeism.

Reply

Attach

Cancel Post Reply

Param Rajesh Joshi (https://northeastern.instructure.com/courses/170748/users/224235)

Feb 15, 2024

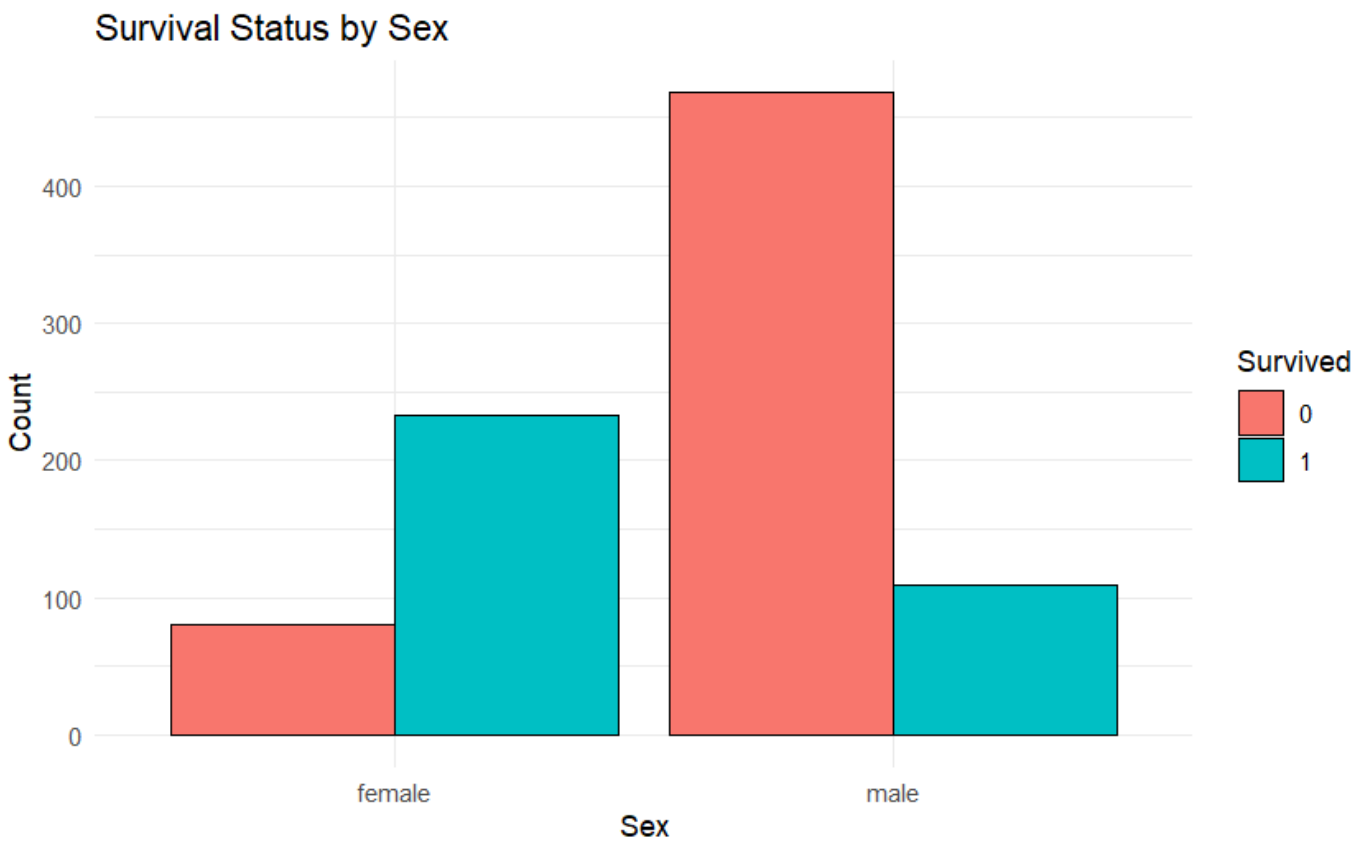
Flash Paper 1. Dataset Description: The dataset chosen for this exploratory analysis is the "Titanic:

Flash Paper

1. **Dataset Description**: The dataset chosen for this exploratory analysis is the "**Titanic: Machine Learning from Disaster**" dataset from Kaggle. This dataset comprises information about passengers aboard the Titanic, including whether they survived or not. It's sourced from Kaggle's Titanic competition, aimed at predicting survival outcomes based on passenger attributes. I opted for this dataset due to its historical significance and the opportunity it presents for analyzing factors influencing survival rates during the Titanic disaster. My aim was to explore the relationship between passenger attributes such as age, sex, and class, and their survival status.

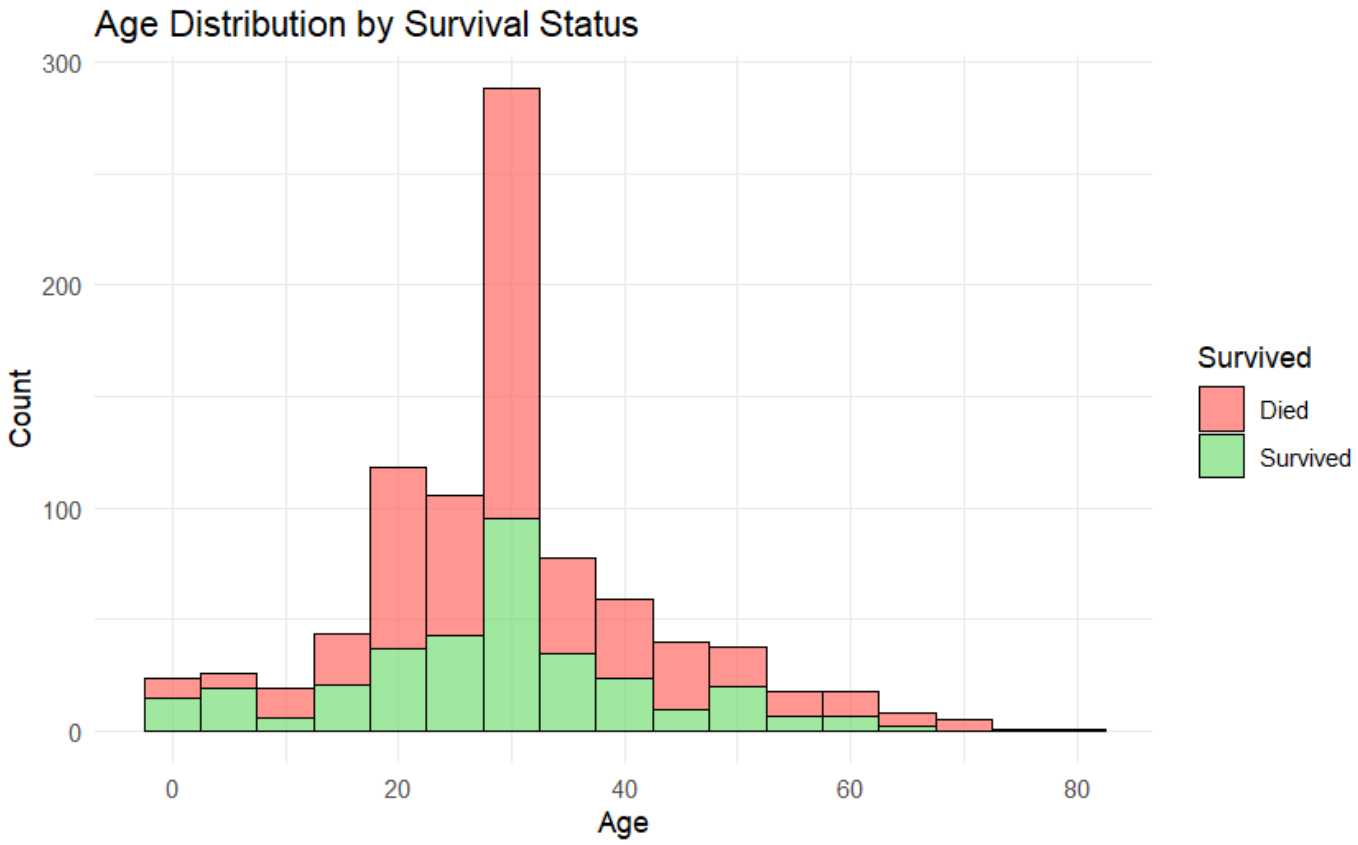
2. Dataset Structure and Preprocessing: The dataset consists of 891 observations and 12 attributes including passenger class, sex, age, siblings/spouses aboard, parents/children aboard, fare, and embarkation port. Upon initial inspection, missing values were found in the "Age", "Cabin", and "Embarked" columns. To address this, I removed the "Cabin" column and imputed missing values in "Age" with the median age, and in "Embarked" with the most common value. Categorical variables were converted to factors for analysis.

3. Exploratory Visualization 1: A bar plot was created to visualize the distribution of survival status based on passenger sex. This visualization revealed potential gender disparities in survival rates, with a higher proportion of females surviving compared to males. This finding aligns with historical accounts of "women and children first" during the Titanic evacuation. Further exploration of survival rates based on other attributes such as passenger class and age could provide deeper insights into the demographics of survivors. Overall, this dataset offers a rich opportunity for analysis, shedding light on the tragic events surrounding the Titanic disaster.



This bar plot illustrates the distribution of survival status (survived or not) based on passenger sex. The x-axis represents sex (male or female), while the y-axis represents the count of passengers. The plot highlights the higher proportion of female survivors compared to males, suggesting potential gender disparities in survival rates during the Titanic disaster.

4. Exploratory Visualization 2: One interesting visualization I created was a histogram showing the distribution of passenger ages. The histogram revealed that the majority of passengers were in their twenties and thirties, with a smaller number of children and elderly individuals. This insight could potentially inform analyses of survival rates based on age groups. Additionally, exploring the relationship between age and survival status could provide valuable insights into the demographics of survivors. Overall, this dataset offers numerous opportunities for exploration and analysis, shedding light on the tragic events surrounding the Titanic disaster.



This histogram visualizes the distribution of passenger ages aboard the Titanic. The x-axis represents age intervals, and the y-axis represents the frequency of passengers within each age interval. The histogram highlights the age distribution of Titanic passengers, showing peaks in the twenties and thirties age ranges. This visualization serves as a starting point for deeper analysis of survival rates based on age demographics. The age distribution of passengers varies widely, with a peak around young adults. There is a notable proportion of children among passengers. The visualization suggests that the survival rate for children might be higher compared to other age groups, as indicated by the relatively larger proportion of green bars (indicating survival) in the lower age range. Conversely, there is a relatively larger proportion of red bars (indicating non-survival) in the middle age range, suggesting that adults in this age group might have a lower survival rate. Overall, this visualization provides insights into the age distribution of passengers aboard the Titanic and how it relates to their survival status.

Reply

Attach

Cancel Post Reply

Keegan Veazey (https://northeastern.instructure.com/courses/170748/users/279956)

Feb 15, 2024

Flash Paper - OkCupid Dataset With experience in exploring social media outlets such as Twitter a

Flash Paper - OkCupid Dataset

With experience in exploring social media outlets such as Twitter and YouTube through previous research, I used this opportunity to check out another popular social media avenue – dating apps. In my search I came across an OkCupid user dataset found here: https://www.kaggle.com/datasets/andrewmvd/okcupid-profiles/data (https://www.kaggle.com/datasets/andrewmvd/okcupid-profiles/data) . This dataset has 59946 rows and 31 features. Each row represents an OkCupid user's profile and includes columns about user demographics (sex, orientation, language spoken, age, etc.), time last active online, as well as textual responses to prompts about hobbies and user experiences. For HW2, I focused on demographic and time-related information since we have not covered language processing in R yet (though this is also an area of interest).

Because I hoped to answer questions around user demographics and their time spent online, I first narrowed the data column-wise – dropping the textual prompt answers and keeping the following columns: sex, orientation, speaks, and last_online. I then checked for NA values (I loaded empty strings "" as NA when reading in the CSV) and found the speaks column (where values are comma-separated lists of languages spoken by the user in a single string form) had 40 NA values. Given the context of the data, I decided to impute the NA values with the number one since it is safe to assume users speak at least one language. I could have also imputed with the average number of languages, though it is notable that the average rounded down to 1 anyway. Next, I created a new column based on the number of commas in the speaks column string (note: num commas + 1 = user language count) using the str_count from the stringr R package. I then moved on to the last_online column, whose values were strings in the form "yyyy-mm-dd-hh-mm" (ex: "2012-06-28-20-30"). Because I was interested in answering the question "What time of day were people most active", I converted the values to datetimes, extracted the hour, and then generated a new column called "hour_active" to store the user's last hour active.

Before getting into the results, note that for the purposes of this assignment and for result

interpretation I assume this data is a representative sample of users on OkCupid.

The most interesting visualization and result I found was about the time of day users were most active (see visual below and sex distribution for data representation context).

sex_distribubution.png (https://northeastern.instructure.com/users/279956/files/26279418?wrap=1&verifier=EQ1gfbMr16gR9fli9mLG3KluS3bmrJHLXve6Xtn9) (https://northeastern.instructure.com/users/279956/files/26279418/download?verifier=EQ1gfbMr16gR9fli9mLG3KluS3bmrJHLXve6Xtn9&download_frd=1)

hour_last_active_by_sex.png (https://northeastern.instructure.com/users/279956/files/26279425?wrap=1&verifier=TrbSsaiOb4w98l4XKSMZNLkGT6KnOjJIlwWs6nA6) (https://northeastern.instructure.com/users/279956/files/26279425/download?verifier=TrbSsaiOb4w98l4XKSMZNLkGT6KnOjJIlwWs6nA6&download_frd=1)

What do these visualizations tell us? From the first graph, we can see that there are more men (60% of users) compared to female users - so the majority of users on OkCupid are men. Also, the Hour Last Active bar chart shows that the hours last active across men and women follow very similar patterns – the distribution of each overall shape is very similar between the two. What's more, we can see that users are least active between 2:00 am - 7:00 am UTC. Activity increases after 7:00 am with most activity between 8:00 pm - 12:00 am (midnight). This finding makes sense since many people work during the day and sleep at night. Why are these interesting visuals? They provide insight about the people using OkCupid and opportunities for growth (e.g: the company may want to find ways to get more women on the app and/or could do more promotions or matching events during late hours).

Edited by Keegan Veazey (https://northeastern.instructure.com/courses/170748/users/279956) on Feb 15 at 12:44pm

Reply

Attach

Cancel Post Reply

-

VINAYAKA H K (*He/Him*) (https://northeastern.instructure.com/courses/170748/users/

186433)

Feb 15, 2024

Motivation and about Data Set: I have chosen The Consumer Behaviour and Shopping Habits Data

Motivation and about Data Set:

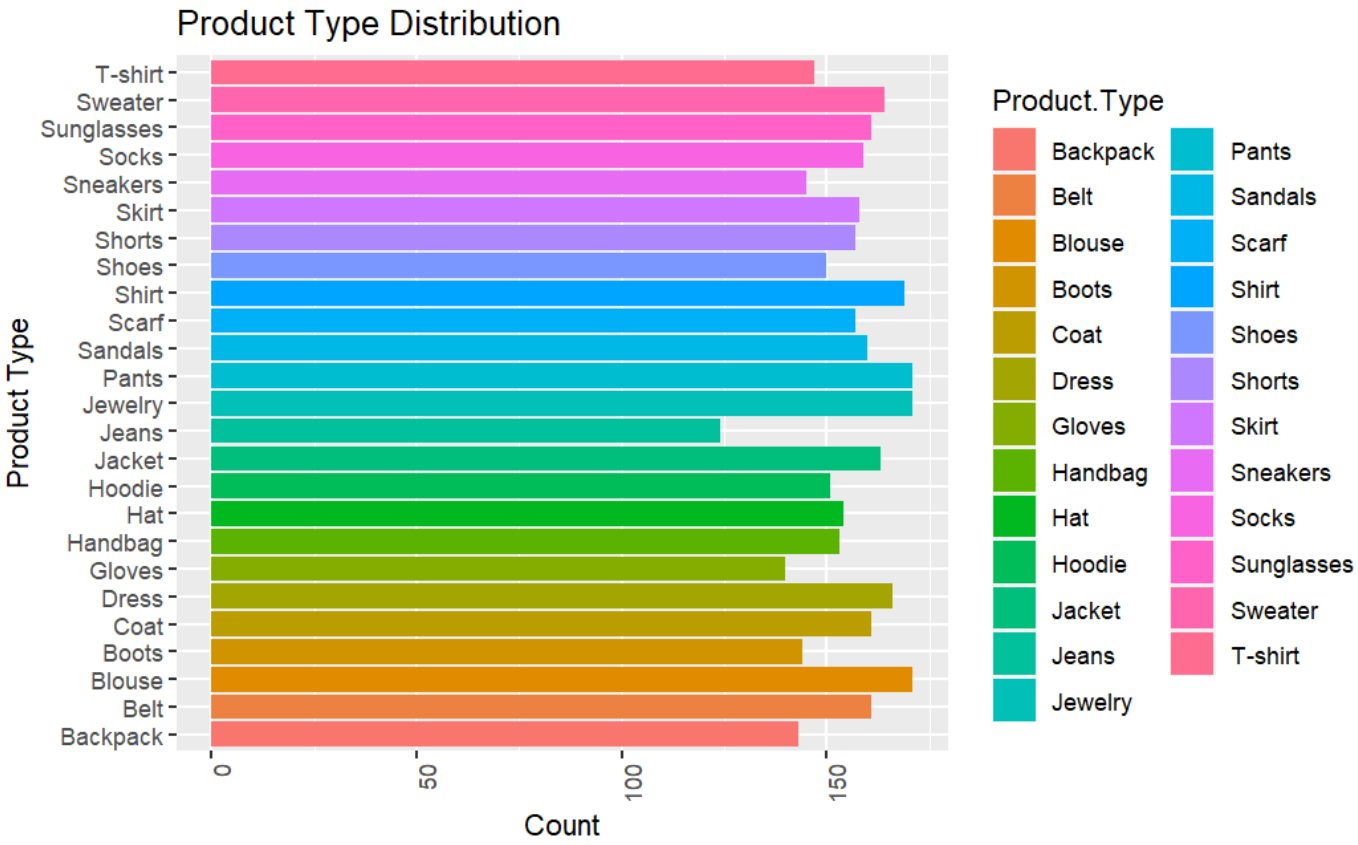
I have chosen The Consumer Behaviour and Shopping Habits Dataset. It provides comprehensive insights into consumers' preferences, tendencies, and patterns during their shopping experiences. I found it in Kaggle (https://www.kaggle.com/datasets/zeesolver/consumer-behavior-and-shopping-habits-dataset?resource=download).I chose this dataset because it contains data that help predict consumer behaviour. Since I work at Co-op, where we are concerned about customer behaviour, I wanted to analyse how customers behave across seasons, what categories of products they prefer, the mode of payment they favour, and when they shop during the year, including which seasons. Based on these data, we can maintain a larger stock of products in inventory. We can also target specific age groups and come up with targeted campaigns, ads, and promotions.

DATA DESCIPITION AND PREPROCESSING:

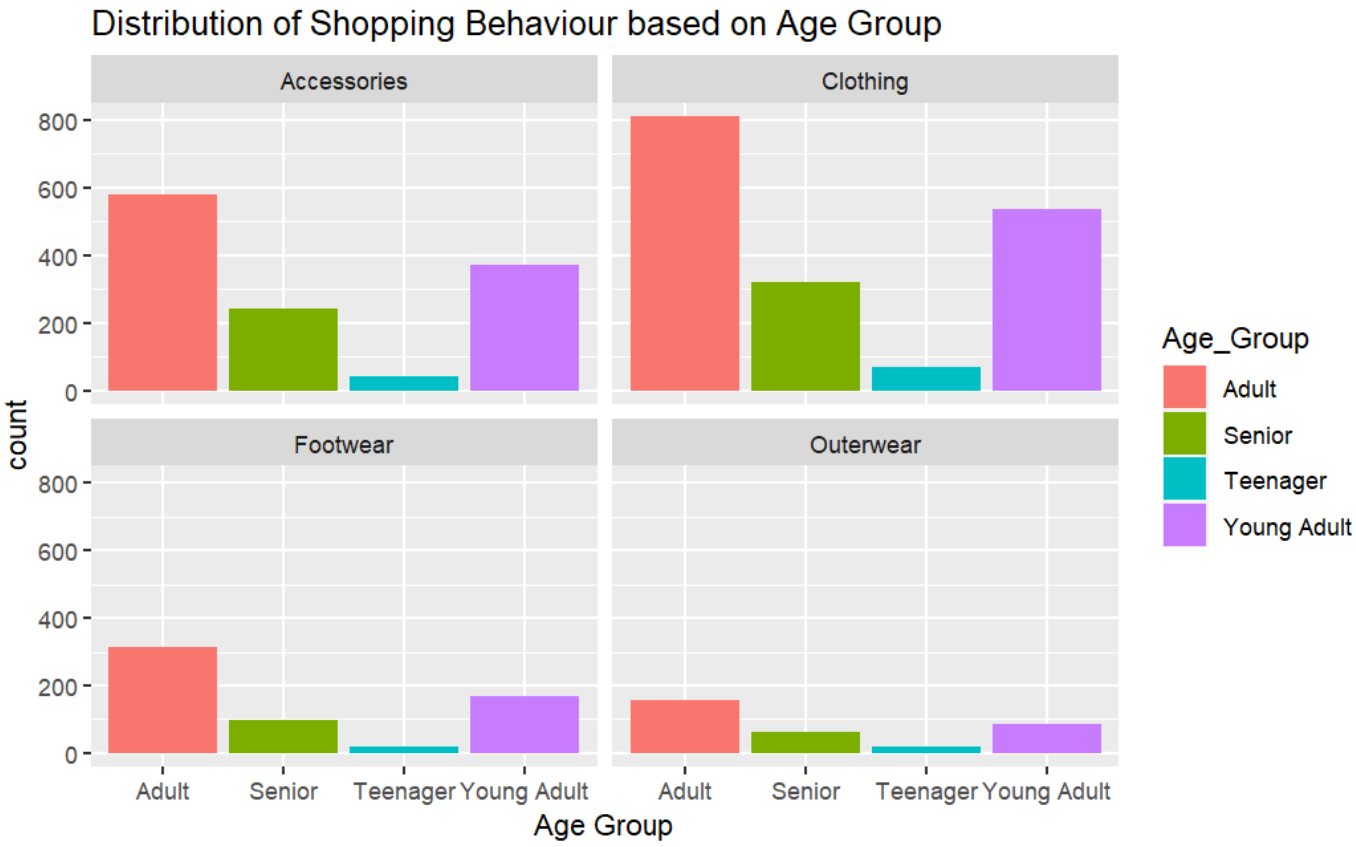
It has following columns Customer ID, Age, Gender, Item Purchased, Category, Purchase Amount (USD), Location, Size, Colour, Season, Review Rating, Subscription Status, Shipping Type, Discount Applied, Promo Code Used, Previous Purchases, Payment Method, Frequency of Purchase. I have checked for NA values. Converted into data frame. Dropped size, colour since, it was not much of help for my analysis. I have renamed 2 columns "Rating" from "Review. Rating" and “Product. Type" from "Item. Purchased”.

ANALYSIS:

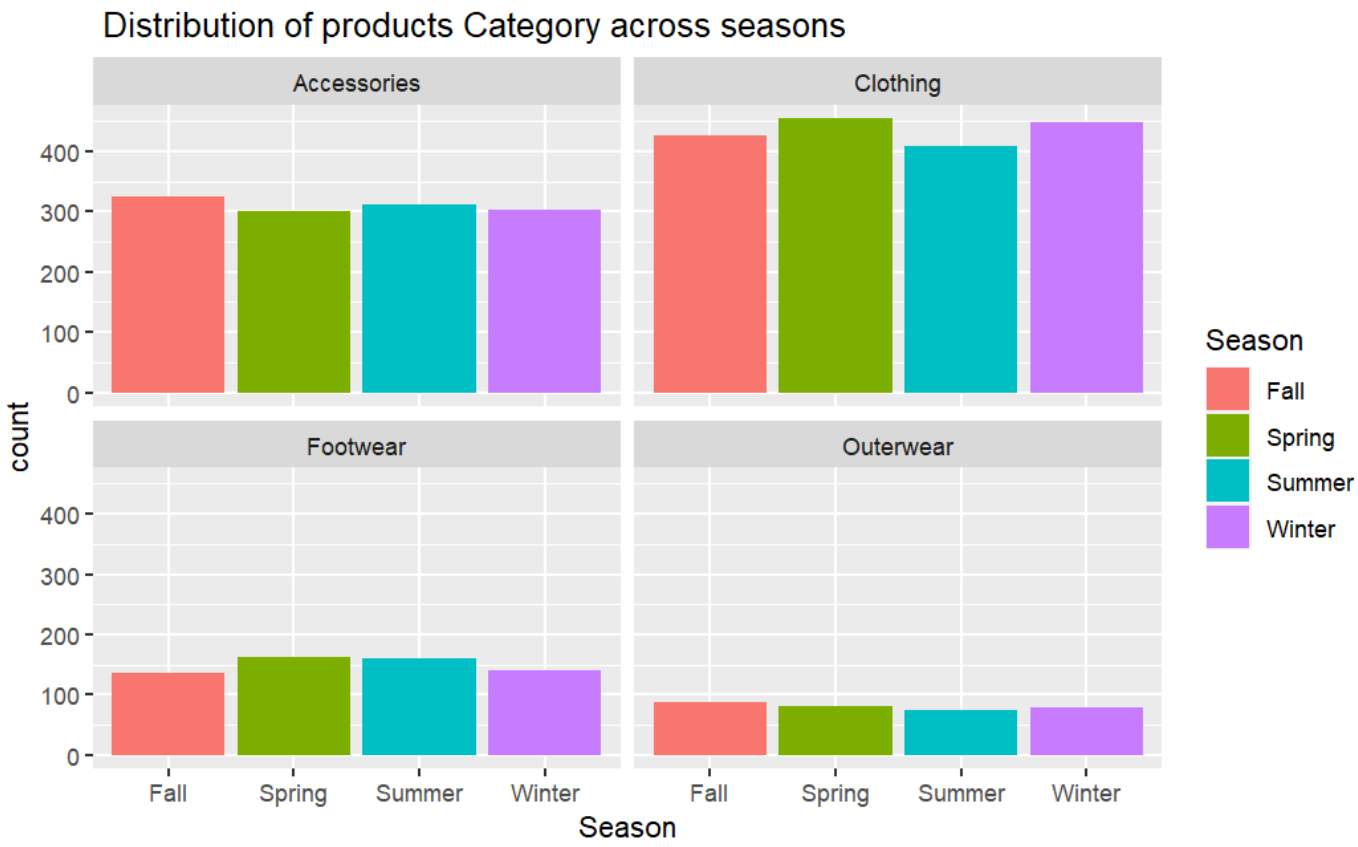
Since I wanted to find out which Product was highly shopped, I plotted the graph to see same. From the below graph we see that the most purchased Items are Boots, Pants, Jewellery. Jeans is the least Purchased Item.



Another Interesting behaviour was to see how people across various age group shop do, who shops more and what products do they shop more and least. From the below graph we see that the Adults and followed by young Adults have purchased the most while Teenager purchased the least. Across all the age group everyone has purchased more in Clothing and least in Outwear.



Another Interesting behaviour was to see how people shop across various seasons of the year so that we can keep those products more in inventory for that season. From the Below graph we see that Customers have shopped for Accessories more in Fall, Clothing in Spring, Footwear in sparing &summer. Outerwear in winter and Fall. And across all the seasons we see clothing has been bought most.



Reply

Attach

Cancel Post Reply

-

Jennisha Christina Martin (*She/Her*) (https://

northeastern.instructure.com/courses/170748/users/306147)

Feb 15, 2024

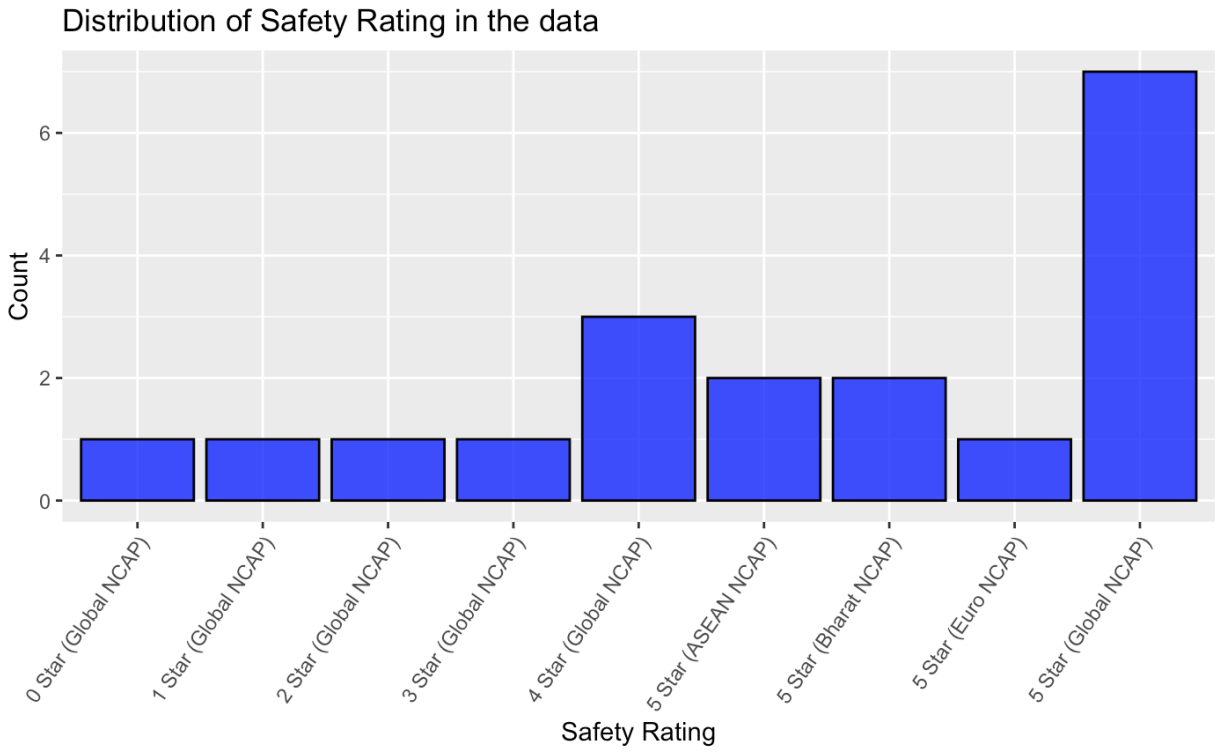
Hi Everyone,Hope you all are doing well.I'd like to share some insights about the dataset which I'd

Hi Everyone,

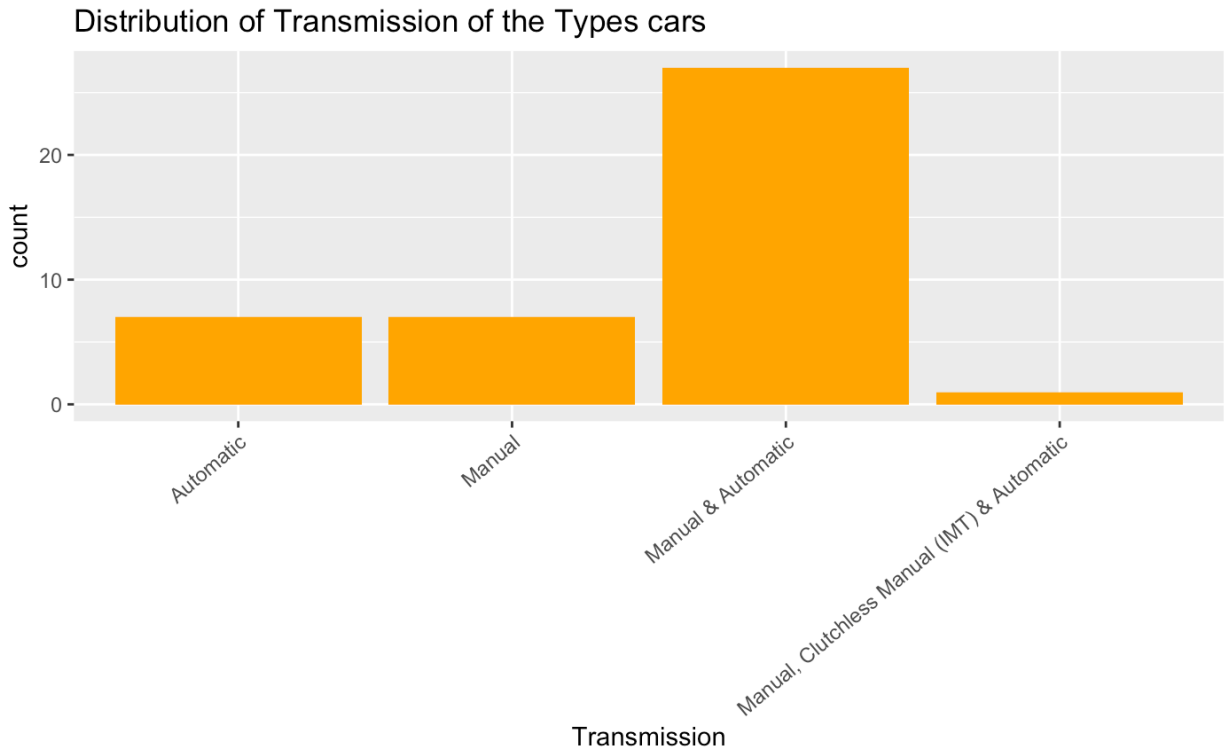
Hope you all are doing well.

I'd like to share some insights about the dataset which I'd used for analysis in my homework 2. The dataset was sourced from Kaggle and can be accessed from the following link: **https://www.kaggle.com/datasets/shiivvvaam/indian-cars-under-20-lakhs?resource=download (https://www.kaggle.com/datasets/shiivvvaam/indian-cars-under-20-lakhs?resource=download)** . This dataset provides an overview of the diverse range of affordable and budget-friendly cars available in the Indian market, focusing specifically on those priced under Rs. 20 Lakhs (equivalent to 24,108.10 USD). I chose this dataset particularly because I found it to be intriguing due to its relevance in providing valuable insights into the world of budget-friendly and feature-rich vehicles in the automotive industry in India, where the car market is diverse and rapidly evolving. Furthermore, my main aim was to delve deeper into the dataset and identify the key trends and characteristics of all the affordable and fuel-efficient cars within this price range, in order to gain invaluable insights into the consumer preferences, market dynamics, and the competitive landscape within the Indian automotive industry. These insights could be beneficial not only for various stakeholders in the industry but also for the consumers to make ready and informed decisions.

This dataset comprises all the details of cars falling within the above-specified price range, including attributes such as the model name, mileage, engine specifications, price, safety ratings, fuel types, transmission options, and seating capacities. Covering a wide spectrum of information, hence making it favourable for analysis. Before proceeding directly with the analysis, I wanted to ensure the dataset was tidy, so I carried out some preprocessing steps to ensure the quality of data is good and suitable for proper analysis. These steps included checking for inconsistencies, null values, and duplicates, and excluding them from the dataset in order to make it unique. Moreover, I created a refined dataset by removing unnecessary attributes with the help of indexing and retained only the essential fields for analysis. Finally, I standardised and refined variable formats by eliminating unnecessary characters which are not allowed in the attribute names.



A visualisation which I found to be fascinating from my analysis is a bar chart depicting the relationship between the cars and their safety ratings. This visualisation provides a detailed overview of the safety ratings distribution within the dataset, offering key insights into the number of cars falling under each safety rating category. A higher star rating here indicates better safety performance of the car. Interestingly, from the visualisation, I observed that there were many cars falling under the higher safety rating category, while considerably fewer of them were having lower ratings, showing a positive trend and indicating that the majority of the cars within the specified price range tend to have better safety performance. Moreover, this demonstrates how the Indian automotive industry prioritises the safety standards in the cars made affordable, in order to meet the evolving needs of the buyers. As safety is a critical factor for many car buyers, this could benefit the consumers by enabling them to compare and understand the safety performance of different vehicles, thereby, making it easier for them to make effortless decisions when purchasing a car, as per their needs.



Another visualisation which drew my interest is the above, which illustrates the correlation between the number of cars and their transmission options, ranging from traditional manual transmission to an innovative combination of manual and automatic technologies. Surprisingly, in this visualisation, I noticed that there were more cars with both manual and automatic transmission compared to the cars with other transmission options. Indicating a strong demand in the market for such affordable vehicles that offer flexibility in driving. Furthermore, the presence of combined transmission option such as clutchless, manual and automatic, is pretty impressive, as it shows the technological diversity within the affordable car ranges in the Indian automotive industry.

Overall, while exploring the dataset, I delved into several key aspects to gain a comprehensive understanding of the affordable cars in the Indian market, present in the dataset. Firstly, I analysed the trends in fuel efficiency among the different cars present within the given price range. This provided insights into buyers' preferences for fuel-efficient vehicles and the technological advancements driving improvements across the Indian car market. Secondly, I analysed the transmission options for the cars available in the dataset, and lastly, explored the safety ratings and the seating capacities of the cars present. Exploring this dataset was truly an eye opener for me, In doing so, I acquired insights into the important factors influencing the trends in the affordable car segments of the automotive industry in India.

Reply

Attach

Cancel Post Reply

-

Arjun Pesaru (https://northeastern.instructure.com/courses/170748/users/311376)

Feb 15, 2024

Dataset-Electric Vehicle Population Data 1) The dataset I chose originates from the Washington St.

Dataset-Electric Vehicle Population Data (http://catalog.data.gov/dataset/electric-vehicle-population-data)

1) The dataset I chose originates from the Washington State Department of Licensing (DOL) and provides information on **Battery Electric Vehicles (BEVs)** and **Plug-in Hybrid Electric Vehicles (PHEVs)** registered in the state. It includes details such as vehicle make, Electric vehicle type, model, year, and possibly registration location.

I chose this dataset for:

- **Relevance**: Electric vehicles (EVs) are becoming increasingly important in efforts to reduce greenhouse gas emissions and combat climate change. Understanding their adoption and distribution can offer insights into the progress of sustainable transportation initiatives.
- **Interest**: The topic of electric vehicles is of significant interest to policymakers, environmentalists, and the general public. Having access to data on EV registration can provide valuable information for various stakeholders.
- **Analysis Potential**: The dataset offers opportunities for analysis and visualization to explore trends in EV adoption over time, geographical distribution of registrations, popular EV models, and potentially even correlations with factors such as demographics or incentives.

Questions I wanted to explore in visualization:

Trend Analysis: How has the number of registered BEVs and PHEVs changed over time in Washington state? Are there any noticeable trends or patterns?

Popular Models: Which BEV and PHEV models are most commonly registered in Washington state? Are there any preferences or trends in vehicle choice?

2) The dataset structure includes various variables related to electric vehicles, such as location,

eligibility for clean fuel incentives, technical specifications like electric range and MSRP, and distinctions between Battery Electric Vehicles (BEVs) and Plug-in Hybrid Electric Vehicles (PHEVs).

Preprocessing steps involve:

- **Loading necessary libraries**
- **Reading the dataset into R**
- **Removing empty spaces and NA values**: We clean up the dataset by eliminating rows with missing data and trimming leading/trailing spaces from all columns.
- **Dividing the "Electric.Vehicle.Type" column**: We split this column into two new ones, "Plug-in Hybrid Electric Vehicle (PHEV)" and "Battery Electric Vehicle (BEV)", based on specific strings within the original column.
- **Removing unnecessary columns**: We streamline the dataframe by excluding irrelevant columns like "2020.Census.Tract", "Electric.Vehicle.Type", etc.

These steps ensure data cleanliness and prepare the dataset for further analysis.

3) Exploratory analysis of the data

Figure-1 The dataset from reveals a notable trend: the prevalence of Battery Electric Vehicles (BEVs) exceeds that of Plug-in Hybrid Electric Vehicles (PHEVs). This suggests a higher adoption rate or preference for BEVs among vehicle owners within the state. Such a pattern could be influenced by factors such as environmental concerns, technological advancements, infrastructure availability, or government incentives favoring BEVs.

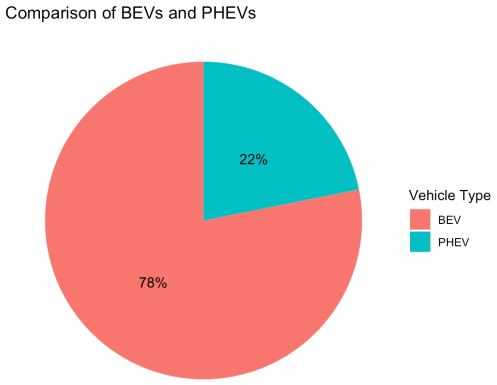


Figure 2-The visualization of the dataset underscores the dominance of Tesla within the Battery Electric Vehicle (BEV) category, with a substantial portion of registered BEVs attributed to Tesla models.This could be attributed to several factors, including Tesla's early entry into the electric vehicle market, its strong brand reputation for innovation and performance, and a diverse lineup of appealing electric vehicle models.

In contrast, Chevrolet emerges as a frontrunner in the Plug-in Hybrid Electric Vehicle (PHEV) category. This leadership position may stem from Chevrolet's efforts to offer affordable and versatile PHEV options. Additionally, these findings may inform future marketing strategies, product development efforts, and policy initiatives aimed at accelerating the transition to sustainable

transportation solutions.

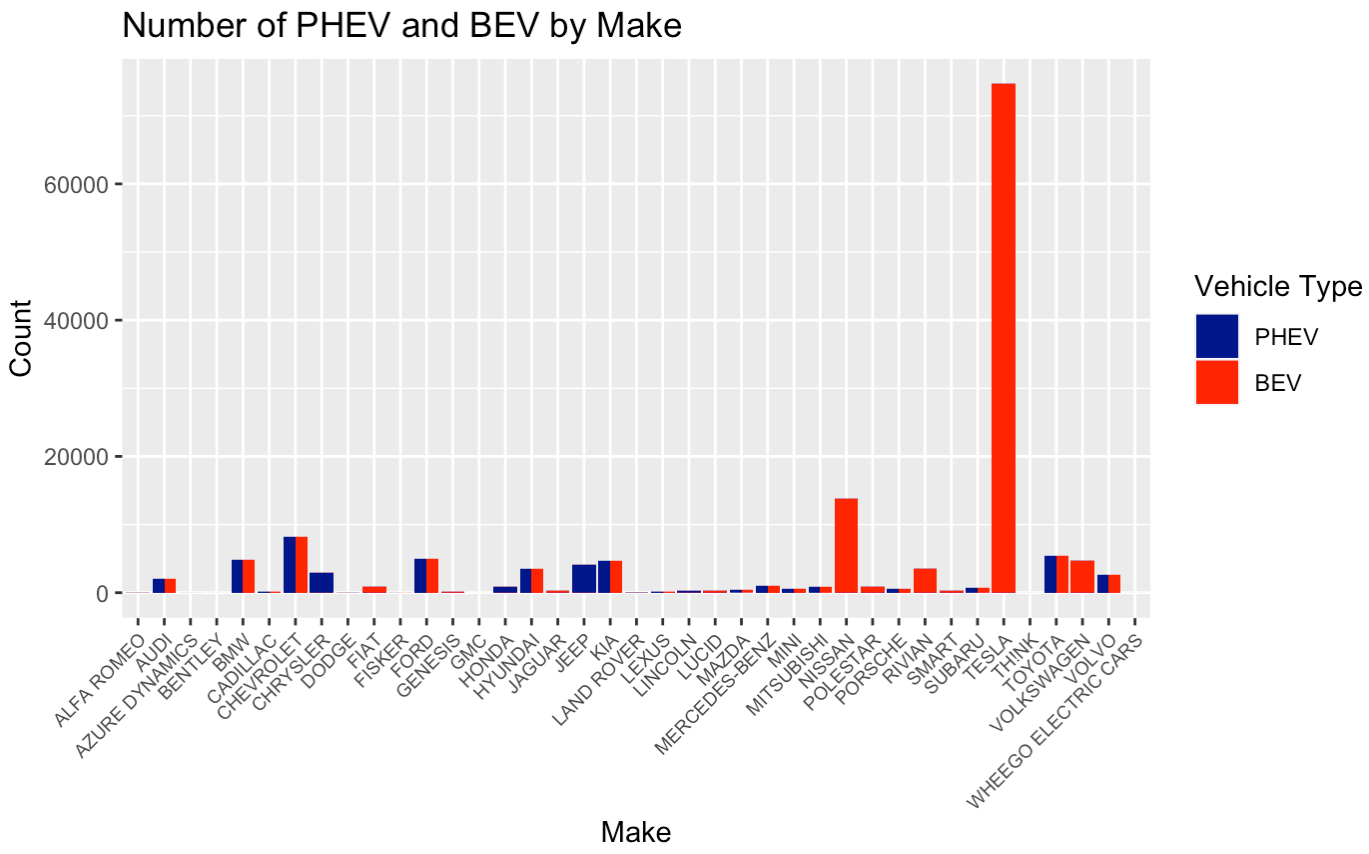
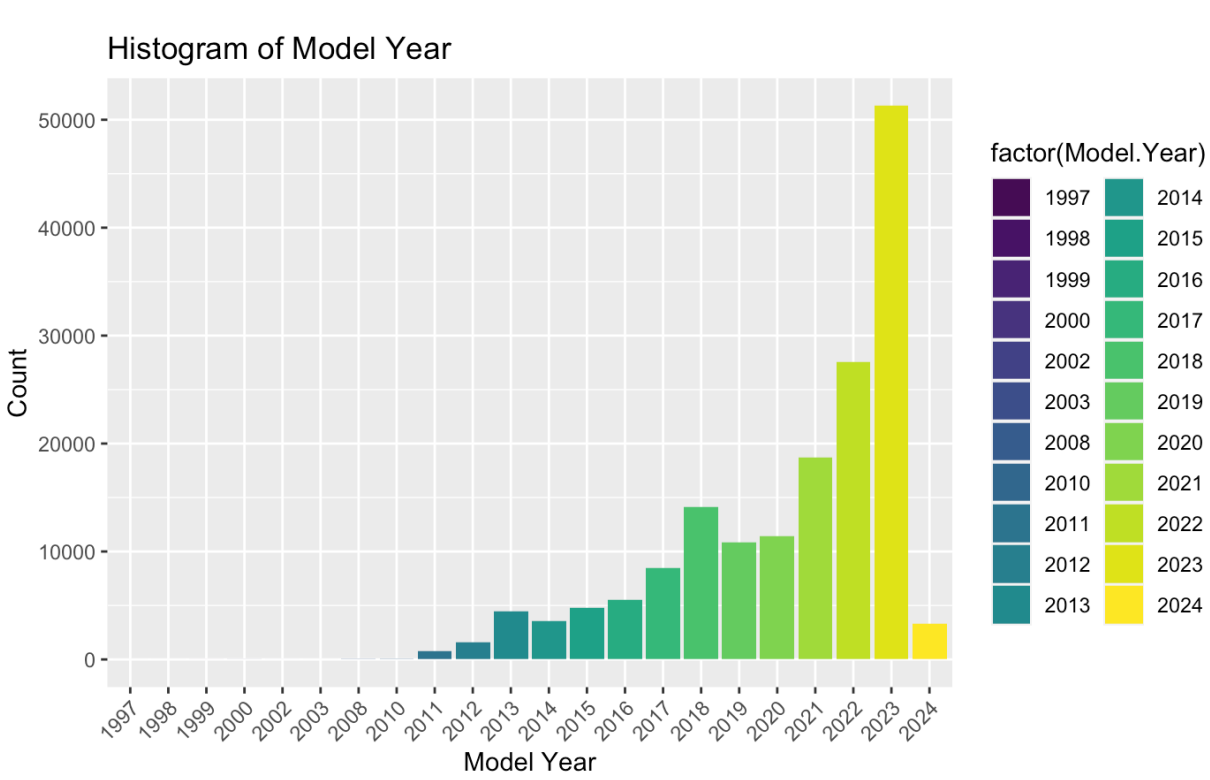


Figure 3-The histogram depicts a noticeable increase in the counts of electric vehicles, particularly with a significant spike observed in the year 2023. The heightened interest and adoption of electric vehicles in 2023 underscore the momentum and acceleration of the shift towards sustainable transportation solutions. This trend aligns with global efforts to mitigate climate change, reduce greenhouse gas emissions, and transition away from fossil fuel-dependent vehicles. Understanding the dynamics behind this spike can inform future strategies for policymakers, manufacturers, and stakeholders in the electric vehicle industry to sustain and further drive the growth of electric mobility.



Overall, The analysis shows a clear preference for Battery Electric Vehicles (BEVs) over Plug-in Hybrid Electric Vehicles (PHEVs), with Tesla dominating the BEV market and Chevrolet leading in PHEVs. A significant spike in EV adoption in 2023 reflects the accelerating shift towards sustainable transportation. Understanding this trend can guide future strategies for promoting electric mobility.

Reply

Attach

Cancel Post Reply

-

Naveen Kavitha Gunasekaran (https://northeastern.instructure.com/courses/170748/users/312978)

Feb 15, 2024

1. The dataset I've chosen is about German-made cars. Here's the source. The dataset includes ne

1. The dataset I've chosen is about German-made cars. Here's the **source.** (https://www.kaggle.com/datasets/ander289386/cars-germany?rvi=1) The dataset includes new and old cars that are for sale. I chose this specific dataset because I am interested in cars as a whole and I need a specific dataset that has less and precise data (like cars from germany) when compared to a generic dataset that has a lot of data from all over the world.
The visualization I did is a comparison between electric vs non electric cars before and after 2015 in Germany. As a result I found that the non electric cars still have high sales compared with electric cars.
2. This dataset has a total of 9 columns. Some of the major ones are OfferType (this describes if the car is used or new), Fuel(I used this for my visualization, this column describes which type of fuel the car uses (Electric, Diesel, Petrol, Other)), make (describes the make of car like BMW, Audi), year(describes the year in which the car was released).
Before processing data, I had to do some pre-processing steps which includes steps like omitting all NA's, removing the duplicates, checking if the dataframe is a tidy.
3. The image I've attached below is the visualization I got as a result. The visualization describes the count of cars categorized by year before 2015 and after 2015 as well as the type of fuel it uses. From the image, we can infer that the usage of cars as a whole peaked after 2015 which increased the count of electric and gas cars.
We can also infer that even though there is a increase in electric cars after 2015, the growth of non electric cars still precedes the growth of electric cars. i.e. The growth of non electric cars after 2015 is still higher than growth of electric cars after 2015. This shows most people still prefer non electric cars over electric cars.

Probelm2_result.png (https://northeastern.instructure.com/files/26281313/download?download_frd=1&verifier=il6iOu3Zla8fObBhyapX80touDWpLF6eqR7PyTqB)

Reply

Attach

Cancel Post Reply

Priyadharshan Sengutuvan (https://northeastern.instructure.com/courses/170748/users/312979)

Feb 15, 2024

The Dataset I took is a Bike dataset which has the detailed data entities that are available to descri

The Dataset I took is a Bike dataset which has the detailed data entities that are available to describe the bike's brand, price, and it's specifications, Which I took it from an openly available dataset in Kaggle, Source (https://storage.googleapis.com/kagglesdsdata/datasets/2839189/4895991/bikesCleaned.csv)

I chose this dataset because I like bikes!? And also it is available for free. I wanted to visualize the change of price in bikes with respect to displacement and also wanted to know if changes in specification would alter the speed and price of the bike. Like if the tire size increases will it increase the speed of the bike.

The dataset contains the bike's manufacturer and has its prices listed down. This has around 45 data columns which discuss the types of transmission, no of cylinders, type of clutch, its displacement and so on which can describe the bikes specifications. I was interested in variables which describe its tires - tire pressure, size etc.. and its displacement which determines its engine compatibility and also price to compare with above mentioned columns. The dataset has some duplicates which has to be deprecated to be able to visualize the data, so used Pivot_longer to make the dataset more tidier so that it can be visualized

The figure I visualized is the one with relation between the price and the displacement of the bikes

The plot shows that the price of the bikes is directly affected by its displacement as the graph shows the price has a positive growth as the displacement of the bike increases

34d9a441-9a5e-4eb1-b9e3-2704f27ad22c.png (https://northeastern.instructure.com/files/26281314/download?download_frd=1&verifier=qZqQqPZgKTVjL0wJpu8lPousSEo4h2BzAZoa9bWJ)

Reply

Attach

Cancel Post Reply

-

Madhuri Krishnamurthy (https://northeastern.instructure.com/courses/170748/users/144410)

Feb 15, 2024

The dataset I used is from Kaggle and is titled "FIFA 21 Messy Raw Dataset for Cleaning & Explori

The dataset I used is from Kaggle and is titled "FIFA 21 Messy Raw Dataset for Cleaning & Exploring."[1] The dataset contains information about FIFA 21 player attributes, such as their ratings, positions, nationality, and various in-game statistics. I chose this dataset because it provides a comprehensive overview of player attributes in the popular FIFA video game series, which is of interest to both gamers and soccer enthusiasts. Additionally, I wanted to explore questions related to player ratings, such as which countries produce the highest-rated players or which positions tend to have the highest ratings.

The dataset includes various columns such as player ID, name, age, nationality, club, position, and different player attributes like height, weight, club, preferred foot, best position, and overall rating. Before visualization, the dataset needed preprocessing to handle missing values, standardize data types, and possibly handle outliers. Steps for preprocessing includes handling missing data by imputation or removal, converting categorical variables into dummy variables if necessary, and scaling numerical variables. Some of the preprocessing steps performed were:

1. We can standardize the names to follow a consistent format using "clean_names" . For example below, the column names 'LongName", "↓OVA", "playerUrl" and "BOV" have been standardised to “long_name”, "ova", "player_url" and “bov” in the data frame. This lowercase naming convention is consistent with the other column names.

2. Select only required columns. The original dataset consisted of 77 columns, out of which I chose 10 columns which were required for visualization.

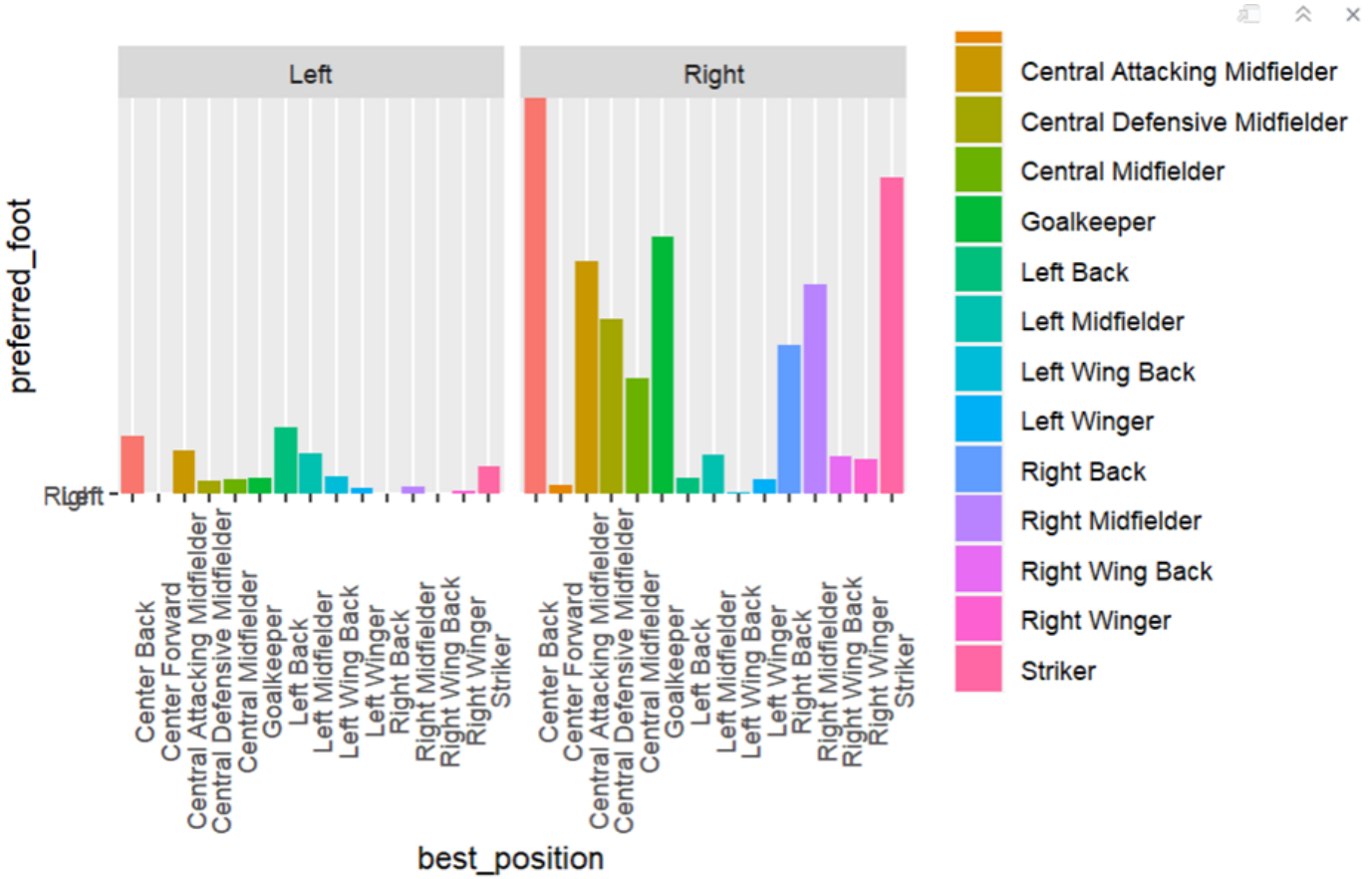
3. Removing ‘newline' characters using gsub: Club column consists of unnecessary ‘newline’ characters.

club <chr>	club <chr>
\n\n\n\nFC Barcelona	FC Barcelona
\n\n\n\nJuventus	Juventus
\n\n\n\nAtlético Madrid	Atlético Madrid
\n\n\n\nManchester City	Manchester City
\n\n\n\nParis Saint-Germain	Paris Saint-Germain
\n\n\n\nFC Bayern München	FC Bayern München
\n\n\n\nLiverpool	Liverpool
\n\n\n\nLiverpool	Liverpool
\n\n\n\nParis Saint-Germain	Paris Saint-Germain
\n\n\n\nFC Barcelona	FC Barcelona

4. "best_position" column consists of short forms, convert the short forms to long forms for better understanding of data. For example,

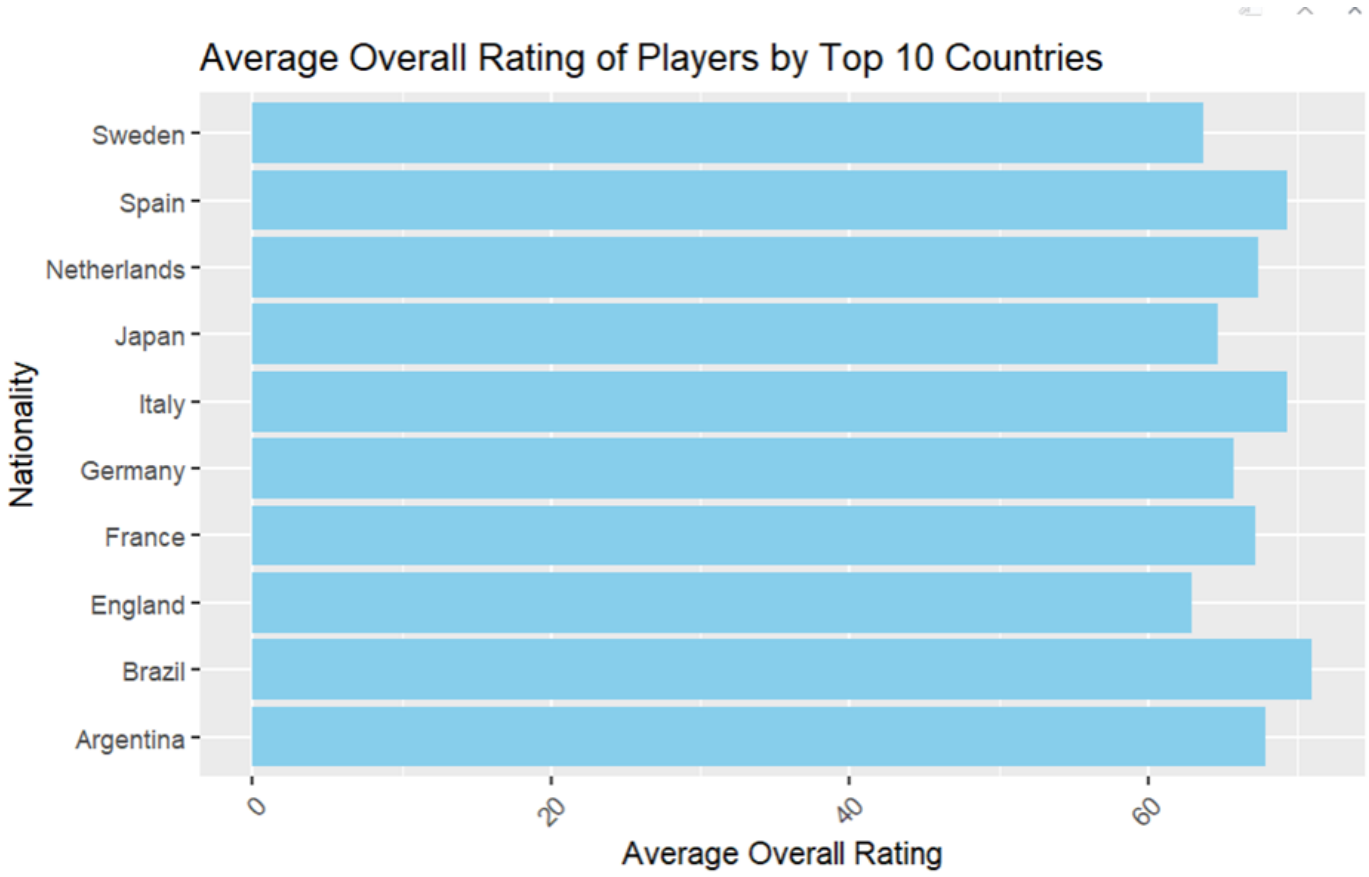
best_position <chr>	best_position <chr>
RW	Right Winger
ST	Striker
GK	Goalkeeper
CAM	Central Attacking Midfielder
LW	Left Winger
ST	Striker
RW	Right Winger
GK	Goalkeeper
ST	Striker
GK	Goalkeeper

The visualization I wanted to see was "What foot is the most common foot to kick with in each position?". I analyzed the dataset to determine the distribution of preferred kicking foot (left or right) among players in each position on the field. The visualization illustrates the relationship between best positions on the field and their preferred kicking foot. The bar chart shows the distribution of players across different positions based on whether they prefer to kick with their left foot or their right foot. Additionally, the chart is divided into separate panels for each preferred kicking foot, allowing for easier comparison between the positions within each category.



From the visualization we can see that most players best kick is with their right foot in Center Back position, followed by the Striker position and Left Wing Back being the least. The best position for left-foot kickers is Left Back position.

Another interesting visualization could be a bar chart showing the average overall rating of players from the top 10 countries with the highest number of players represented in the dataset. This visualization would provide insights into which countries produce the highest-rated players on average in FIFA 21. After preprocessing the dataset, I computed the average overall rating for players from each country and then selected the top 10 countries by player count. The resulting bar chart showed the average overall rating for players from these countries.



From this visualization we can see that Brazil is the country with highest overall rating of players.

[1] https://www.kaggle.com/datasets/yagunnersya/fifa-21-messy-raw-dataset-for-cleaning-exploring?resource=download (https://www.kaggle.com/datasets/yagunnersya/fifa-21-messy-raw-dataset-for-cleaning-exploring?resource=download)

Reply

Attach

Cancel Post Reply

-

Melissa Pax (https://northeastern.instructure.com/courses/170748/users/216259)

Feb 15, 2024

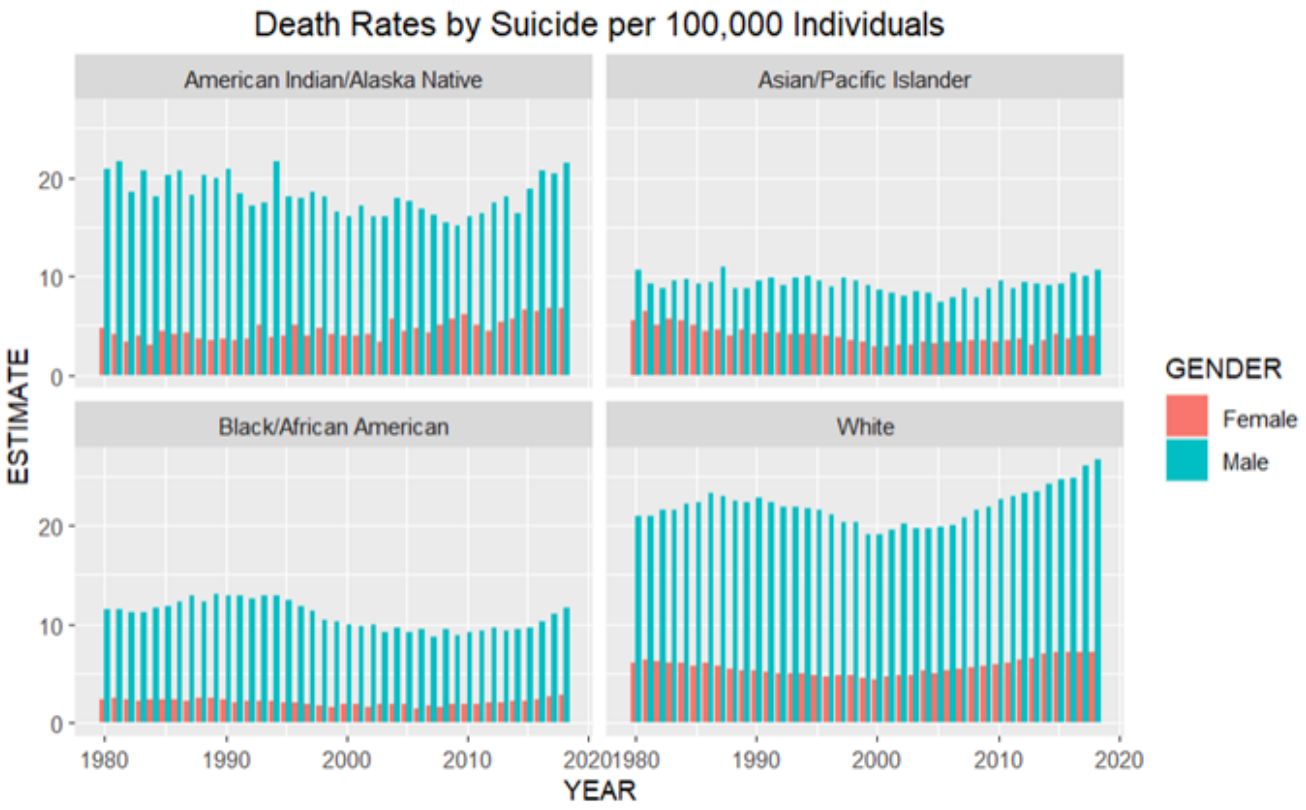
Mental health conditions have well-researched disparities across race and gender. Unfortunately, th

Mental health conditions have well-researched disparities across race and gender. Unfortunately, the outcome of some of these conditions is suicide. This assignment picked the dataset “Death rates for suicide, by sex, race, Hispanic origin, and age: United States” from the National Center for Health Statistics to better understand the fatality rates caused by suicide across different demographics [1]. Visually, the author wanted to visualize the differences in race and gender in suicide death rates to gain a better understanding of these outcome gaps.

The dataset looked at death rates for suicide by age, gender, race, and Hispanic/Latino identification between the years 1980 and 2020. Intersectionality of these demographics was included, meaning they would analyze different combinations of age, gender, race, and Hispanic/Latino identification to assess how rates change when considering different aspects of identity. Each analysis used either crude or age-adjusted death rates per 100,000 individuals in a population.

Many variables of the dataset were different identification keys for the study, which were not particularly useful. The ones proving to be more useful were the following: UNIT, STUB_LABEL, STUB_NAME_NUM, YEAR, AGE, and ESTIMATE. UNIT determined if the rate calculated was age-adjusted or crude, YEAR labeled the year when the death rate was calculated, AGE labeled the age group, and ESTIMATE was the estimated number of deaths attributed to suicide. STUB_LABEL was a numeric value that labeled the analysis based on which demographics (either age, race, gender, or Hispanic/Latino identity) were accounted for in the analysis, while STUB_NAME gave exact character labels to the analysis (e.g. female, white or male, Asian). This was a messy organization, so separate columns were created for gender, race, and Hispanic/Latino identity and extracted the information from STUB_NAME. After this extraction, the UNIT was edited to only include the labels of “age-adjusted” or “crude,” as they were previously also labeled with “Death rate per 100,000 individuals,” which was redundant information.

Below is a graph looking at race and gender of death rates:



This figure was interesting to me because the graph highlights how much larger the death rate for men was compared to women across all races. While some gaps were larger than others, the pattern remained the same, which I thought was interesting. Another interesting observation was how much larger the death rates were for white and American Indian/Alaska Native populations compared to Asian/Pacific Islander and Black/African American populations. I also found it interesting how death rates appeared to be increasing in most recent years, especially for white and American Indian/ Alaska Native populations. This data exploration left me with more questions than answers; I think comparing this dataset with another dataset looking at non-lethal suicide attempts would also be interesting and may provide further clues as to why there are differences between different populations. Additionally, exploring potential reasons for the slight increase in suicide rates between white and American Indian/Alaska Native populations would probably be a worthwhile research endeavor.

Citation

[1] National Center for Health Statistics. Death rates for suicide, by sex, race, Hispanic origin, and age: United States. Data accessed [Last accessed date]. Available from https://data.cdc.gov/d/9j2v-jamp.

Reply

Attach

Cancel Post Reply

-

Qihui Fan (https://northeastern.instructure.com/courses/170748/users/269175)

Feb 15, 2024

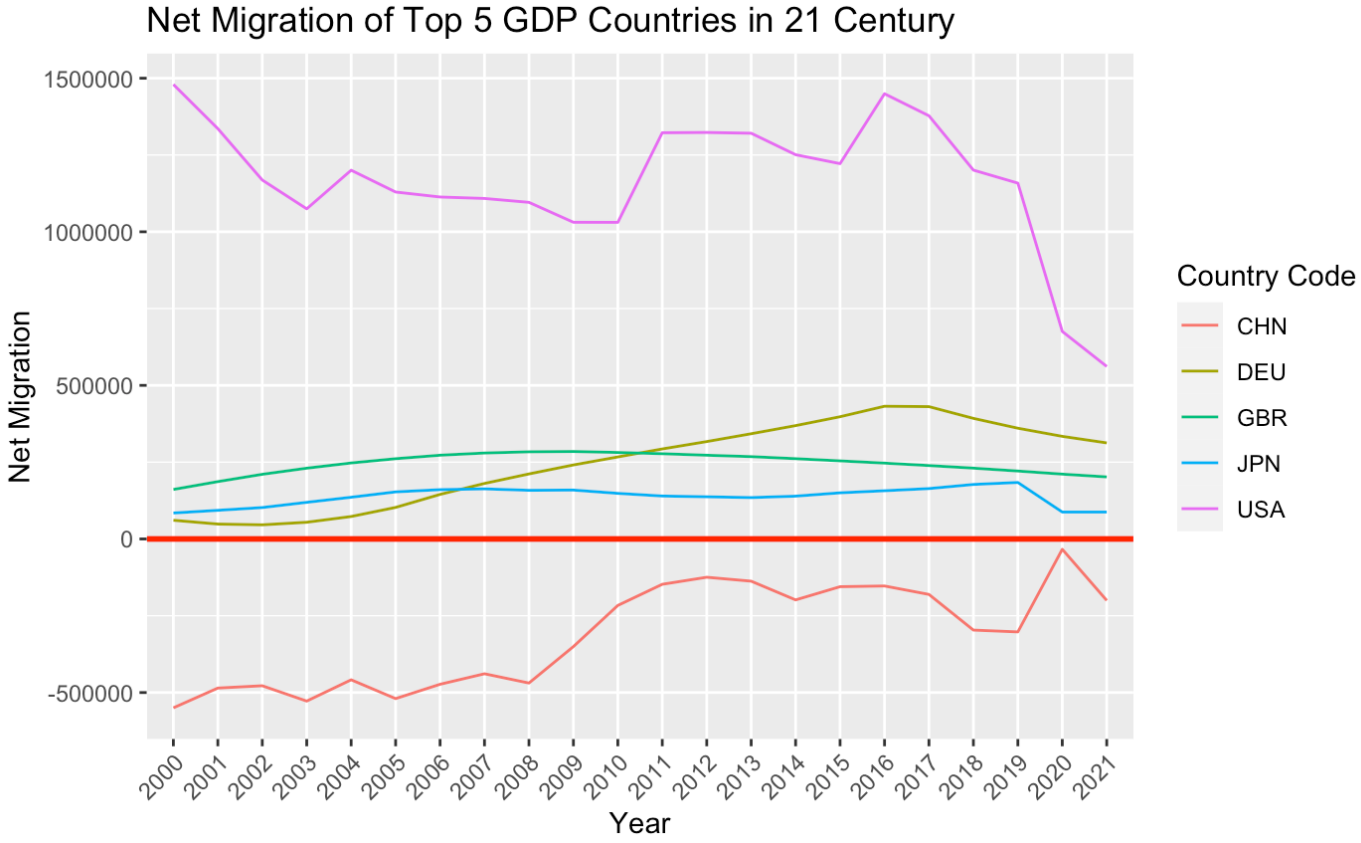
The dataset I chose is from the World Bank about the net migration of different countries from 1960

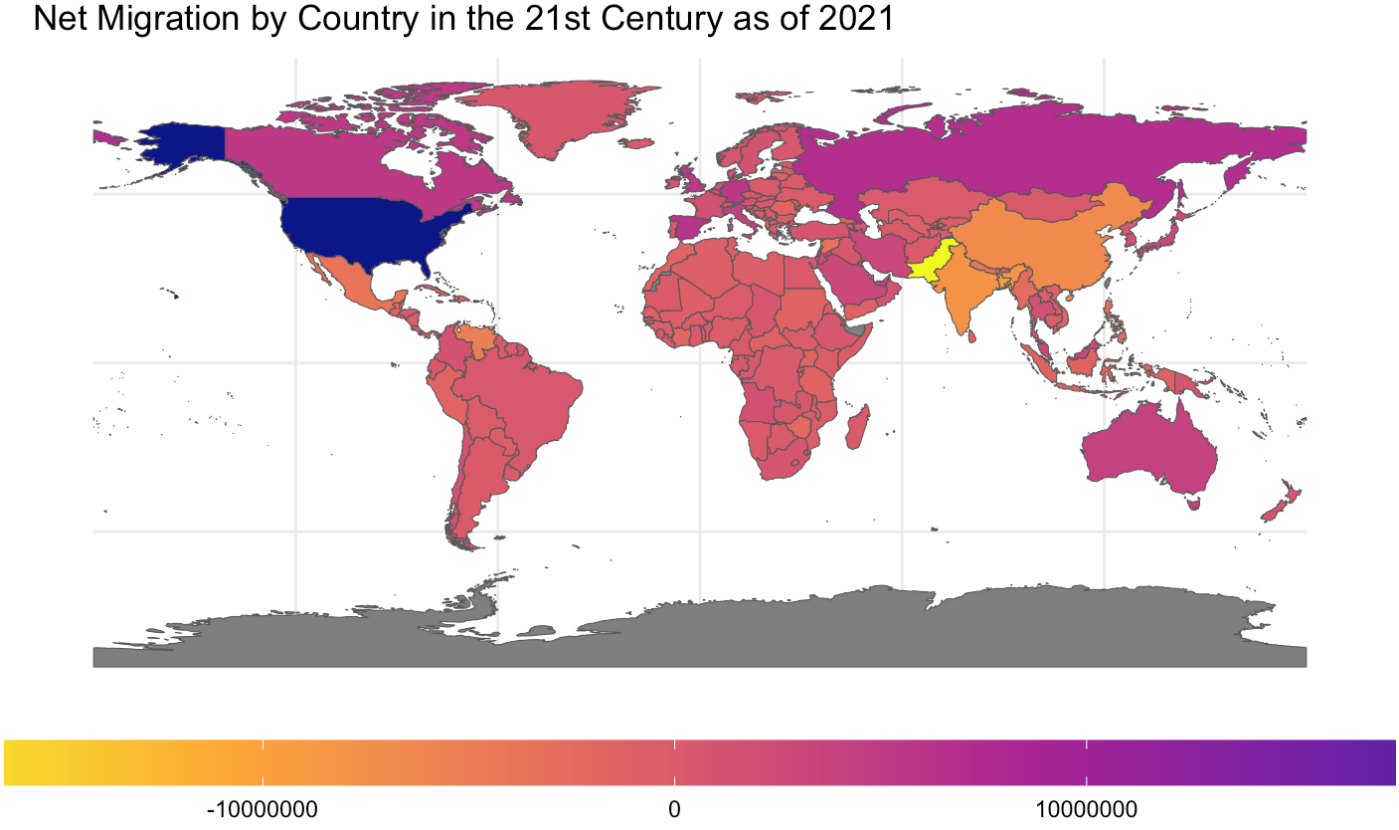
The dataset I chose is from the World Bank about the net migration of different countries from 1960 to 2022, where by definition, “net migration is the net total of migrants during the period, that is, the number of immigrants minus the number of emigrants, including both citizens and noncitizens”. The dataset is from United Nations Population Division, World Population Prospects: 2022 Revision, and is available at https://data.worldbank.org/indicator/SM.POP.NETM?name_desc=false (https://data.worldbank.org/indicator/SM.POP.NETM?name_desc=false) . I am interested in this dataset because I want to see how all kinds of historical events such as different wars, pandemic, political instability would have an impact on migration in a more quantitative side of view, globally. In addition, trying to build a relationship between the economic development of a country with its net migration in a period of time could also be interesting.

The dataset has columns Country Name, Country Code, Indicator Name, Indicator Code and year 1960 to 2022 where each column of the year has the number of net migration of listed countries. Country names and country codes are char data type. The country codes are made with 3 characters for each country in iso_3 standard.

The problem of this data set is also obvious from my above description, in column names, a column named by empty string "" is observed. Observations are grouped by each recognized individual countries, political unions such as European Union, geographic regions such as Asia, and country income group. In these observations, 2022 data are all missing values. So the empty string column "" located at the last of the column is removed. The column of 2022 is also removed since this column only contains null observation values. Indicator Name and Indicator Code are irrelevant to our analysis so they are also removed, only left country code, country name and year 2000 to 2021 columns. Then the next step to consider is if to only consider the net migration by countries for further data analysis, those aggregate data should be removed. From another dataset that comes with the

main dataset, a country code without a value in the region value indicates aggregate data, so aggregate data could be removed this way. Finally, list the net migration data correspondingly to each year as a single observation to form a tidy dataset.





My proposed two plots shows how net migration in top 5 GDP countries changes over 2000 to 2021, and the situation of net migration generally around the world in 21 century(data is only available up to 2022). From the line graph you may observe that China is a significant outlier compared with other top 5 GDP countries, though in fact China has the second largest GDP volume globally. From the world map you may also observe that developing countries are more likely to have negative net migration number which indicates people are immigrating out, while developed countries are more likely to have positive net migration number which indicates people are emigrating in. In addition, due to the inconsistency of some of countries' borders, some regions are marked grey.

Reply

Attach

Cancel Post Reply

-

Gautam Reddy Chandupatla (https://northeastern.instructure.com/courses/170748/users/217199)

Feb 15, 2024

Cervical cancer is a significant public health concern globally, particularly in regions with limited acc

Cervical cancer is a significant public health concern globally, particularly in regions with limited access to screening and healthcare services. Analyzing datasets related to risk factors for cervical cancer can provide valuable insights for healthcare professionals, researchers, and policymakers in several ways.

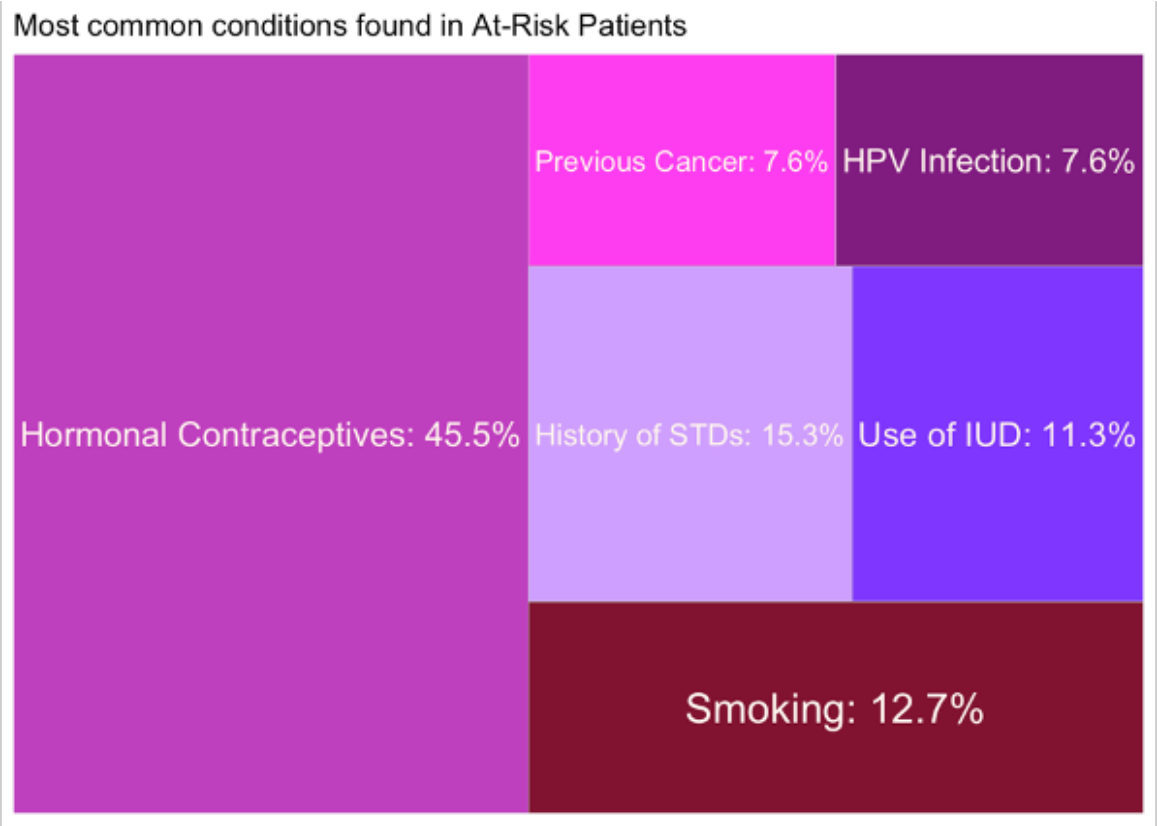
This dataset is obtained from Kaggle (Source:https://www.kaggle.com/datasets/loveall/cervical-cancer-risk-classification (https://www.kaggle.com/datasets/loveall/cervical-cancer-risk-classification)). It contains a list of risk factors for Cervical Cancer and has details about different Biopsy Examinations which are used for detecting Cervical Cancer.

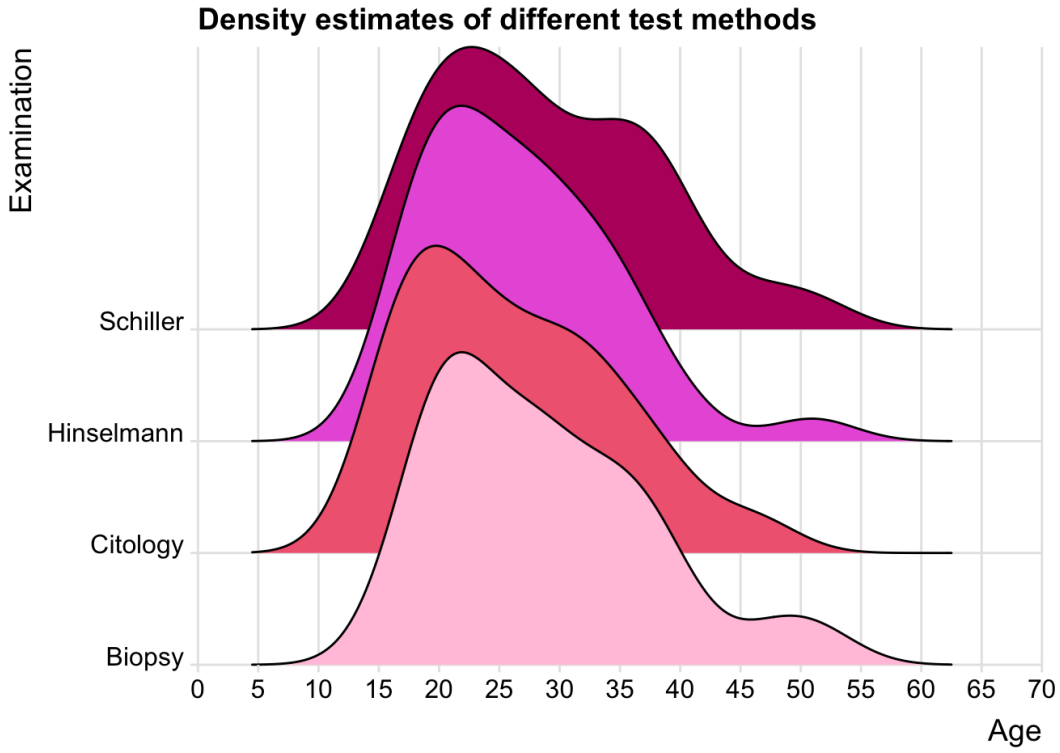
Analyzing this dataset related to risk factors for cervical cancer can contribute to efforts to reduce the incidence, morbidity, and mortality associated with this disease, ultimately improving public health outcomes. Some questions which I would like to explore from this dataset are - Is “AGE” of a women

a risk factor for cervical cancer?, Is there a risk involved in having more sexual partners and pregnancies?, Does smoking causes risk of cervical cancer?, Which STDs increase chance of cervical cancer?, How accurate are the prediction results of different medical examinations?

This dataset contains 858 rows and 36 different columns. Factors related to smoking, STDs, sexual partners, pregnancies and different medical examinations are the variables of interest.

Dropped column “STDs..Time.since.first.diagnosis” and “STDs..Time.since.last.diagnosis” since it has 91% missing values. Replaced NA values with mode for all factor columns. Replaced NA values with mean for all numeric columns. Rounded off the mean in numeric columns since no numeric columns can have decimal values. Different examination results are spread across multiple columns with column names encoding the name of the medical examination. We used pivot_longer() to transform these columns into a tidy representation, and then extract the medical test from the column names.





Cervical Cancer has several risk factors and they've been categorized and analyzed. The above Treemap serves the purpose of quantifying the most common risk factors that have been observed in patients at risk of cervical cancer. Hormonal contraceptives being the most common factor found between at risk patients, taking these pills for more than 5-10 years exponentially increases the risk of cervical cancer. Followed by Smoking, use of IUDs and having a history of STDs. Apart from this, there are a variety of tests available for detecting cervical cancer with Biopsy being the best method. However, a person can pick any test that they feel comfortable with as long as they get tested and diagnose cancer in its early stages to seek proper treatment.

Edited by Gautam Reddy Chandupatla (https://northeastern.instructure.com/courses/170748/users/217199) on Feb 15 at 3:31pm

Reply

Attach

Cancel Post Reply

-

Yifan Qiao (https://northeastern.instructure.com/

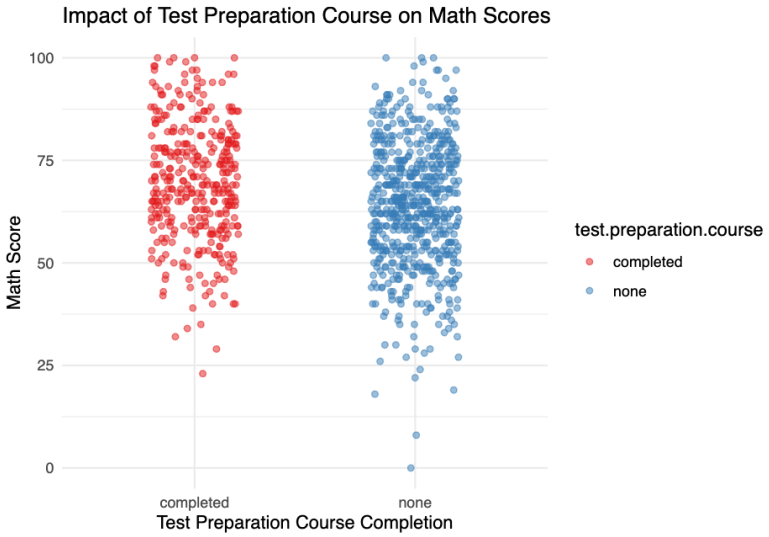
courses/170748/users/154898)

Feb 15, 2024

The dataset utilized in this analysis is called "Student Performance in Exams", which was from Kag

The dataset utilized in this analysis is called "Student Performance in Exams", which was from Kaggle platform. It was contributed by Jakki Seshapanpu and is publicly accessible on Kaggle's website through the link[1]. I chose this dataset because it can show the relevance in understanding factors that influences students' performance between varioussubjects. The dataset provides a comprehensive view of student academic scores and other related factors, like gender, prepare, parents, providing an opportunity to explore relationship between socio-demographic variables and academic achievement.

There are eight variables in this dataset: Gender, Race/Ethnicity, Parental Level of Education, Lunch type, Test Preparation Course completion, Math score, Reading score, and Writing score.
To ensure the data quality, it is important to perform preprocessing steps. At first, I used read.csv() function to import the dataset into R Studio, then I checked the dataset for missing values using sum(is.na(students_data)). Through this process, I identified one missing value and decided to remove the corresponding row using the na.omit() function. Finally, I also verified the data types to ensure the analysis's suitability.



Among the visualization chart (shown above), a important observation arises from the scatter plot. It illustrates the correlation between completing test preparation courses and math scores. Obviously, students who finsihed these test tend to obtain a higher math score, indicating the potential efficacy of such interventions in bolstering academic performance. This shows how important it is to have special classes to help students do better in their studies.

To sum up, my study of the 'Student Performance in Exams' dataset has found some interesting things about how different factors like where you come from affect how well you do in school.The pictures I showed, especially the one with dots, which shows how much better students do in math if they took extra classes, and the ones with boxes that show how boys and girls do differently in school, all tell us that special help can make a big difference in how well students do."

[1] Kaggle. (n.d.). Student Performance in Exams. Retrieved from https://www.kaggle.com/datasets/spscientist/students-performance-in-exams?resource=download

Reply

Attach

Cancel Post Reply

-

Elizabeth Coquillette (*She/Her*) (https://northeastern.instructure.com/courses/170748/users/302731)

Feb 15, 2024

I used a dataset called "Global Trends in Mental Health Disorder: From Schizophrenia to Depressic

I used a dataset called "Global Trends in Mental Health Disorder: From Schizophrenia to Depression" that I got on Kaggle (https://www.kaggle.com/datasets/thedevastator/uncover-global-trends-in-mental-health- disorder?resource=download).

I chose this dataset because I thought it was an interesting topic and would make for interesting visualizations. I did hesitate before using it because it's unlikely that it's reliably accurate, especially in countries or regions in which mental health issues are not widely accepted or diagnosed. The prevalence numbers indicate the percentage of the population who has been identified or diagnosed

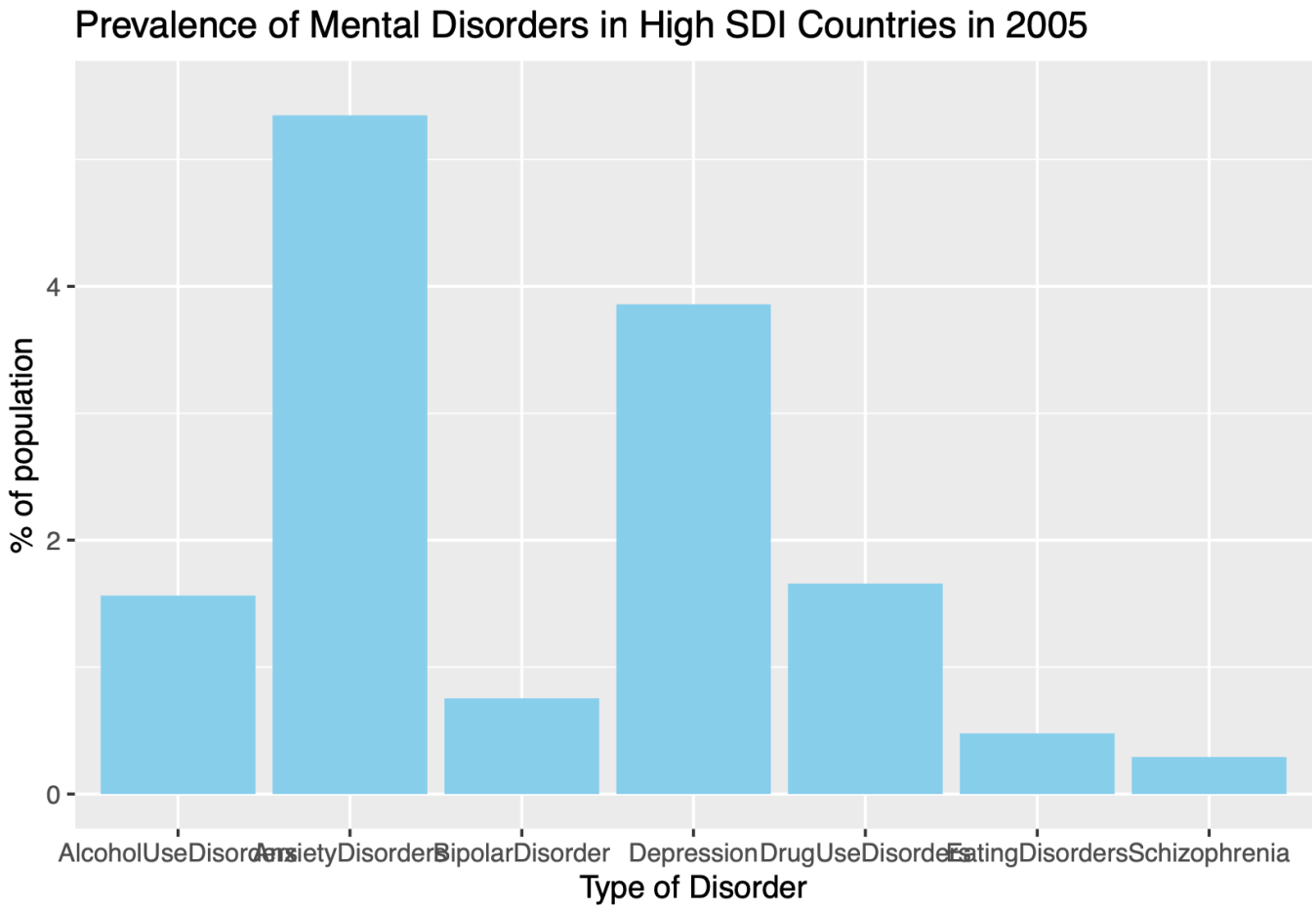
as having that disorder, and in many countries that reported percentage would be lower that the actual prevalence of those symptoms in the population. However, I thought it would still be interesting to compare rates over time.

The dataset covers the prevalence of certain mental health disorders in particular countries/regions and years. Each year of data has a corresponding row that says the "entity" (by which they mean country, region, or category, as I'll describe below), the year, and the prevalence of seven different mental health issues: schizophrenia, bipolar disorder, eating disorders, anxiety disorders, drug use disorders, depression, and alcohol use disorders. The prevalence was listed as percentage of the population with that diagnosis.

The "entity" column contained individual countries, broader regions (e.g. "Central Europe" or "North Africa and Middle East"), income levels (e.g. "High-income" or "High-income Asia Pacific), and SDI levels (e.g. "High SDI" or "Low SDI"). SDI stands for Socio-Demographic Index, which is meant to reflect the overall development level of a country or region based on a combination of income per capita, education level, and total fertility rate.

Not a lot of preprocessing was necessary. The main required step was to change the prevalence values to numeric format using as.numeric(), as in the dataset they were in character format.

Below is a bar graph that I made showing the prevalence of the different types of disorders in high SDI countries in 2005. It shows that anxiety and depression were by far the most prevalent disorders, and both were more than twice as prevalent as the third most prevalent disorder (drug use disorder). Drug use and alcohol use disorders had similar rates, and bipolar disorder, eating disorders, and schizophrenia had significantly lower rates. It would be interesting to make a paired bar graph that included different years and/or different SDI levels, in order to compare the two in that way also.



Reply

Attach

Cancel Post Reply

-

Kuldeep Vishal Choksi (https://northeastern.instructure.com/courses/170748/users/261743)

Feb 15, 2024

Dataset Narrative and Source The dataset titled "2019-2020 Submission MF_LL231_7142020," so.

- **Dataset Narrative and Source**

The dataset titled "2019-2020 Submission MF_LL231_7142020," sourced from **https://catalog.data.gov/dataset/2019-2020-submission-mf-ll231-7142020 (https://catalog.data.gov/dataset/2019-2020-submission-mf-ll231-7142020)** , offers a quantitative narrative on the inclusivity measures within New York City schools, specifically concerning LGBTQGNC students. This dataset piqued my interest due to its potential to reveal the depth of inclusivity efforts in the educational sector. It provides a platform for assessing the impact of Gender Sexuality Alliances and the breadth of training provided to educational staff—a reflection of the commitment to creating a supportive atmosphere for all students.

When delving into the dataset detailing the inclusivity measures of New York City schools, several key questions were at the forefront of the exploration. Primarily, the aim was to understand the extent to which Gender Sexuality Alliances (GSAs) were present and active. Further, there was a keen interest in examining the distribution of LGBTQGNC training hours among various staff categories, to uncover any imbalances that could inform future educational training programs. The exploration sought to check the reflection of such training on the school's overall environment and ethos toward LGBTQGNC students. Additionally, identifying any disparities in training could highlight areas requiring policy reinforcement or strategic focus. Ultimately, the questions aimed to gather insights on how prepared educational institutions are in fostering a supportive and affirming atmosphere for all students, a critical component for holistic development and well-being.

- **Data Preparation and Transformation**

Within the dataset's 920 entries, I encountered a blend of school identifiers and quantitative measures of inclusivity training, which required meticulous cleaning to ensure completeness. Missing values, potentially skewing the analysis, were diligently replaced with zeros. The data's transformation into a long format was a crucial step in facilitating a direct and meaningful comparison of the training hours across different staff roles. This process was not just about tidying up; it was about crafting and preparing a dataset that was ready to show us the important picture.

- **Interpreting the Visual Data**

The visualization "LGBTQGNC Training Hours by Staff Category" was particularly revealing. It showcases the allocation of training hours to different school staff categories, highlighting a commendable concentration on teacher training. This focus suggests a recognition of the pivotal role teachers play in student development and inclusivity. Yet, it also subtly points to the need for a more

inclusive approach to training, as administrative and support staff also play vital roles in shaping a school's culture. The graph serves as both an acknowledgment of the progress made and a prompt for continued dialogue and action toward comprehensive inclusivity training programs.

In essence, this visualization is a confluence of data-driven insights and the human aspects of educational policy—it tells us where we are and gently guides us towards where we need to go to ensure every member of the school community is an active participant in creating a welcoming environment for LGBTQGNC students.

Screenshot 2024-02-15 at 5.00.07 PM.png (https://northeastern.instructure.com/files/26285871/download?download_frd=1&verifier=qN8xtWvWztM02RJDlc6AlJNqlM1icXe4Xah78k7d)

Reply

Attach

Cancel Post Reply

-

Suraj Mishra (https://northeastern.instructure.com/courses/170748/users/270497)

Feb 15, 2024

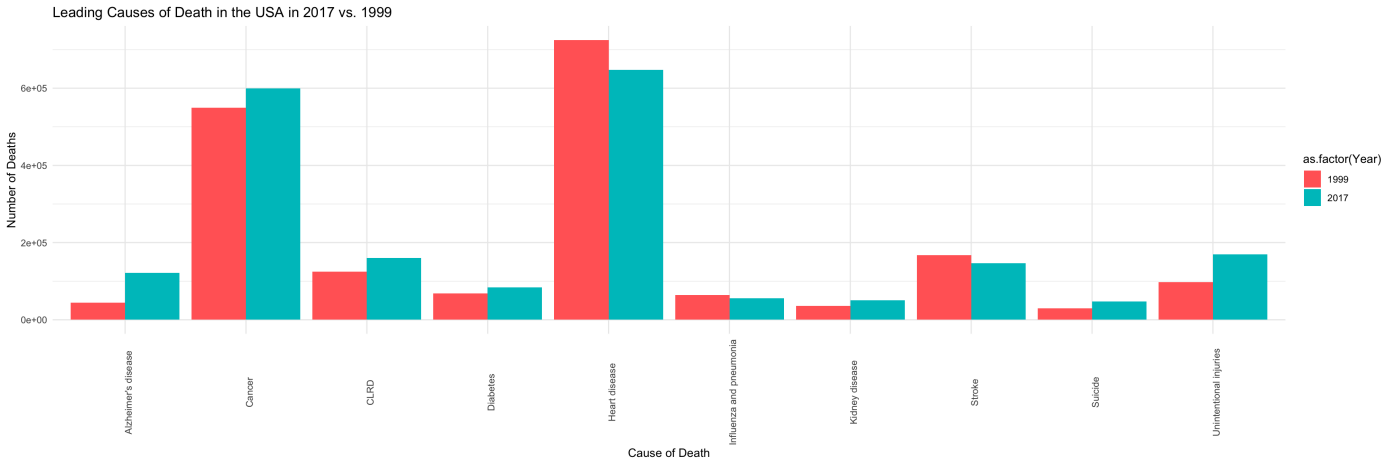
--> I chose the "Leading Causes of Death" dataset from the National Center for Health Statistics (N

--> I chose the "Leading Causes of Death" dataset from the National Center for Health Statistics (NCHS) which is a part of the Centers for Disease Control and Prevention (CDC), available through Data.gov (Citation: National Center for Health Statistics (NCHS) - Leading Causes of Death: United States. Centers for Disease Control and Prevention. Available at: https://catalog.data.gov/dataset (https://catalog.data.gov/dataset)). The reason why I chose this dataset is that it provides an extensive examination of death trends throughout the USA. I was drawn to analyze this dataset to uncover how major causes of death have evolved across the years starting from 1999 up until 2017. With the visualization for the data, I wanted to address the following questions :
a.) How have the leading causes of death evolved from 1999 to 2017?

b.) Has there been a noticeable change in deaths attributed to preventable causes, such as suicide, over the study period?

--> The dataset is structured with variables such as ‘Year’ (which records the year of deaths), ‘Cause_Name’ (which denotes the cause of death), ‘State’ (which indicates the location of recorded deaths), and ‘Deaths’ (which provides the total count).

To render the data suitable for visualization, I executed several preprocessing steps. I initiated the preprocessing by correcting the data type of ‘Year’ to integer from its previous type, double. Subsequently, I modified the character columns: ‘113 Cause Name’, ‘Cause Name’, and ‘State’ to factor for precise analysis. For clarity and easier referencing, I renamed the columns. Furthermore, I filtered out and removed the value “All Causes” from the ‘cause_name’ column to narrow the analysis down to specific causes of death. Following this, I eliminated all the rows with missing (“NA”) values to ensure the completeness of the dataset while maintaining data integrity. Ultimately, I removed all duplicates from the dataset to ensure that it contains only unique records.



--> The bar graph I examined revealed enduring and emerging patterns in the leading causes of death in the USA for the selected years of 1999 and 2017. The visualization indicated that 'Heart Disease' remained the most prevalent cause of death over this period, with 'Cancer' consistently following. While these two diseases have maintained their positions at the top, there has been a slight decline in deaths from 'Heart Disease' in 2017, suggesting a positive impact of health initiatives and awareness. Conversely, the increase in deaths by 'Suicide', although still the least common, necessitates heightened mental health focus. The visual data underscores the need for sustained efforts in combating not only chronic physical illnesses but also addressing mental health care proactively.

Edited by **Suraj Mishra (https://northeastern.instructure.com/courses/170748/users/270497)** on Feb 15 at 6:13pm

Reply

Attach

Cancel Post Reply

Parwaz Sarao (https://northeastern.instructure.com/courses/170748/users/265092)

Feb 15, 2024

1. Describe the dataset and where it comes from (making sure to cite the data source). Explain why

1. Describe the dataset and where it comes from (making sure to cite the data source). Explain why you chose this dataset and what questions you wanted to explore in your visualization.

Answer 1: The dataset in question originates from the Washington State Department of Licensing, from the website https://catalog.data.gov/dataset/electric-vehicle-population-data.This particular dataset provides a registry of electric vehicles (EVs) in Washington State, showcasing various attributes such as Vehicle Identification Number (VIN), county, city of registration, state, postal code, model year, manufacturer (Make), model, type of electric vehicle, eligibility for Clean Alternative Fuel Vehicle (CAFV), Electric Range, Base MSRP, Legislative District, DOL Vehicle ID, Vehicle Location and Electric Utility, 2020 Census Tract.

I chose to focus on this dataset likely due to the increasing relevance and importance of electric vehicles in the context of environmental sustainability and the emerging automotive industry. The dataset tells about the adoption rate and popularity of different EV models and types, which is valuable for understanding market trends, regional preferences, and the progression of EV technology.

The questions that I wanted to explore in this visualization are:

What is the Distribution and count of electric vehicles by manufacturer and model year between 2019 and 2023?
What are the trends in electric vehicle production, particularly for certain manufacturers like Tesla,

Nissan, Kia, BMW, and Chevrolet?
When comparing the year-over-year growth rates, which manufacturer is expanding their presence in the EV market most rapidly?

2. Describe the structure of the dataset and the variables of interest. Describe any preprocessing needs (tidying, cleaning, transformation, etc.) and describe the steps you took to perform the preprocessing.

Answer 2:

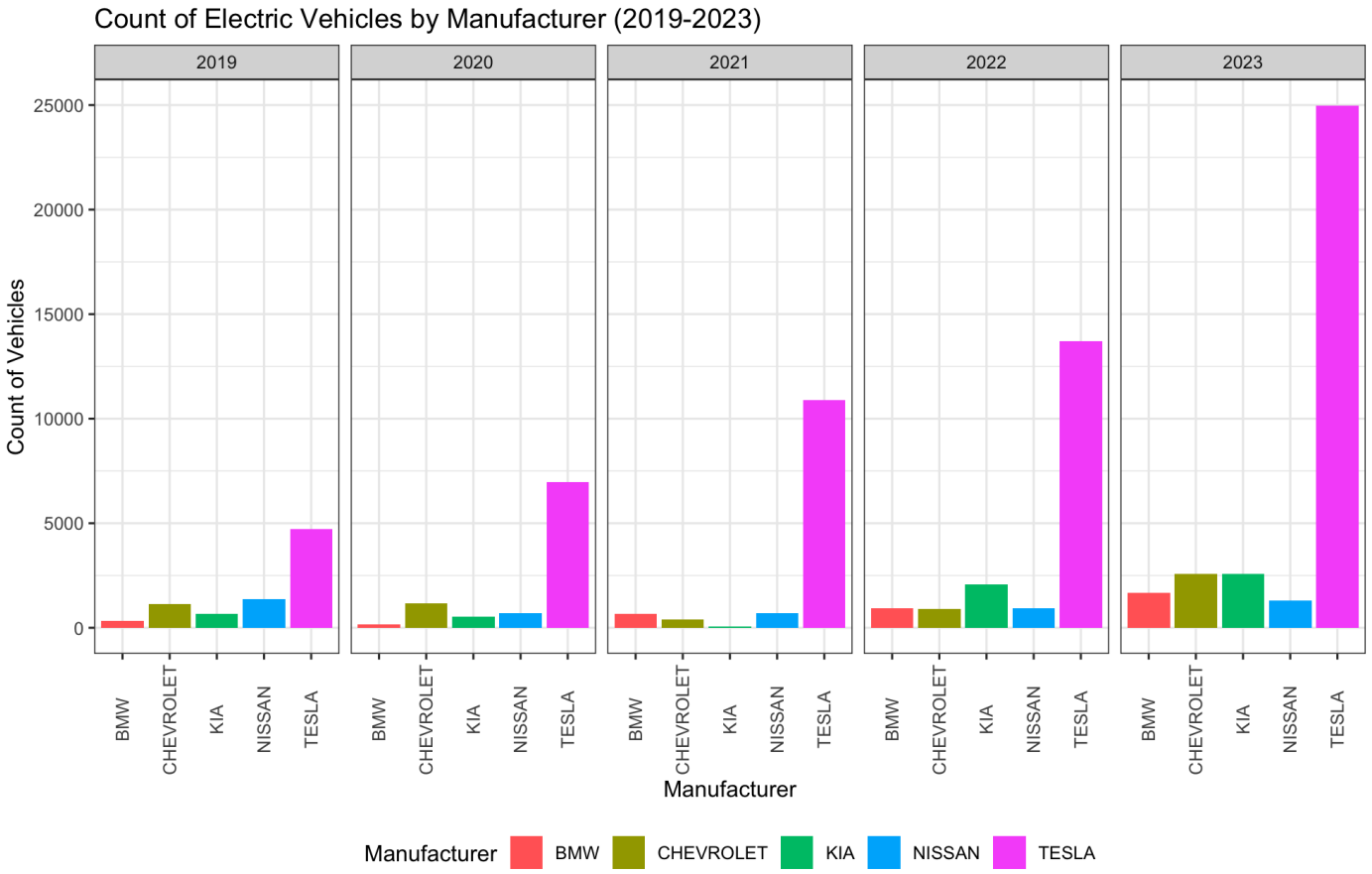
The dataset is structured as a tabular collection of records, each pertaining to an electric vehicle (EV) registered in the state of Washington.The variables of interest are:

1. **VIN (1-10)**: The unique identifier for each vehicle, which is a truncated form of the full Vehicle Identification Number.
2. **County**: The county in the United States where the vehicle is registered.
3. **City**: The city of registration for the vehicle.
4. **State**: The U.S. state where the vehicle is registered, which in this case is Washington (WA).
5. **Postal Code**: The ZIP code for the vehicle's registered address.
6. **Model Year**: The year in which the vehicle model was produced.
7. **Make (Manufacturer)**: The brand of the vehicle.
8. **Model**: The specific model of the vehicle produced by the manufacturer.
9. **Electric Vehicle Type**: The type of electric vehicle, such as Battery Electric Vehicle (BEV) or Plug-in Hybrid Electric Vehicle (PHEV).

There are some preprocessing steps that I performed.Initially, the data was tidied by selecting only the essential columns required for the analysis, which streamlined the dataset by eliminating unnecessary columns.Then after that a cleaning process was undertaken to remove any records containing NA (missing) values. This step was aimed at avoiding any distortions in the findings that missing data might cause. Then after that the 'Make' column was renamed to 'Manufacturer' to enhance understandability, particularly for those who might not be as familiar with automotive industry.

3. Present at least 1 figure that is interesting to you and describe your observations and any key takeaways from the visualization and your exploration of the dataset.

Answer 3:



Observations from the Visualization:

The graphs shows a dramatic and consistent increase in the number of Tesla vehicles increasing each year, culminating in a significant spike in 2023.
In contrast, BMW, Chevrolet, Kia, and Nissan display relatively stable numbers across the years with minor year-to-year variations.
The number of Chevrolet and Kia cars were lowest in 2021.
The number of BMW cars were lowest in 2020.

Key Takeaways:

Tesla's Dominance: The visualization confirms Tesla's dominant position in the EV market, with a strong year-over-year increase in number of cars, possibly reflecting effective marketing, advancements in technology, or broader market shifts towards electric mobility.
Impact of External Factors: The dips for Chevrolet and Kia in 2021, and for BMW in 2020, may point to the impact of external factors on consumer purchasing decisions and the ability of manufacturers to respond to market demands.
Market Growth: Overall, the electric vehicle market shows growth over the years, but this growth is not evenly distributed among manufacturers. Tesla's aggressive growth trajectory highlights its success in capitalizing on increasing consumer interest in EVs.

Reply

Attach

Cancel Post Reply

-

Anurag Akrathuveetil (https://northeastern.instructure.com/courses/170748/users/306209)

Feb 15, 2024

This is visualization of H1b approval and denial data for 2021, 20222, 2023.I chose this dataset be.

This is visualization of H1b approval and denial data for 2021, 20222, 2023.

I chose this dataset because almost all the international students that are studying here, intend to work here and receiving an H1B visa status is of utmost importance for the individual. In the Year 2023 alone there were 483,000 request for H1B visa status out of which only about 17% are approved. This is done through a lottery format by the US Government. Now that lottery isn't in the hands of anyone, getting to the point where an Employer is willing to sponsor you for your visa is a tedious process by itself. This is why I felt this data would be interesting not only for me, but also a lot of my peers. With this data Visualization I intend to gain and provide some statistics or insights that might help people with their H1B journey

Data

The data extracted is the H1B Employer dataset from kaggle
https://www.kaggle.com/datasets/nivedithavudayagiri/h1b-employer-data-2021-2023/data

- Below given are the variables
  1. Fiscal Year
  2. Employer
  3. Initial.Approval
  4. Initial.Denial

5. Continuing.Approval
6. Continuinng.Denial
7. NAICS
8. Tax.ID
9. State
10. City
11. ZIP

Preprocessing/Data transformation

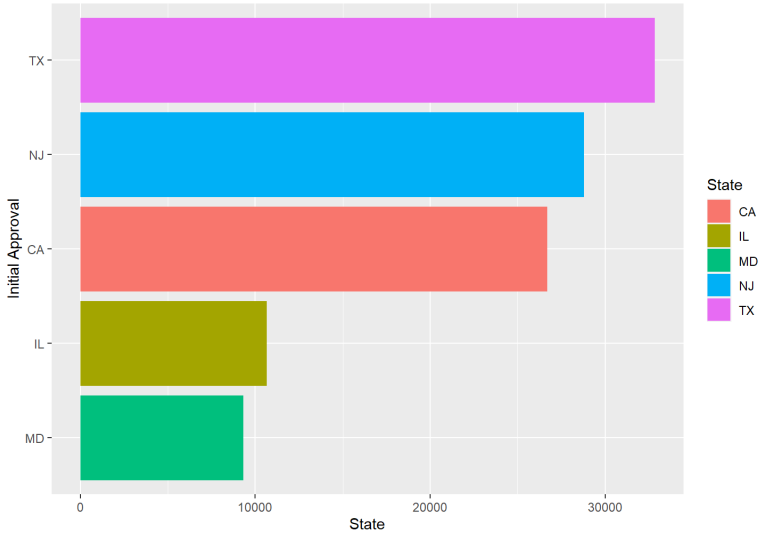
- The dataset is divided into 3 years 2021, 22, 23
  So the first step done was to merge all the data using the {rbind} function.
  Upon reviewing I got 1261 'na' values
- Tax.ID and ZIP has few 'NA' values but we will not be using those columns so discarded those columns from the dataset
- Upon further evaluation it can be seen that "NACIS" column has only one unique value '54' so will discard it too.

For the last Visualization, I wanted the top 3 Employers for the years of 2021, 2022, 2023.

1. First I grouped by Employer. Summarized according to the columns Initial.Approval,Initial.Denial,Continuing.Approval,Continuing.Denial.
2. Then Manually looped through the data for each year and through the columns and added the max value to a new data frame.
3. Upon have the necessary data, I plotted it with geom bar as stat -> Identity and position -> dodge
4. And faceted wrap according to the columns of Initial.Approval,Initial.Denial,Continuing.Approval,Continuing.Denial

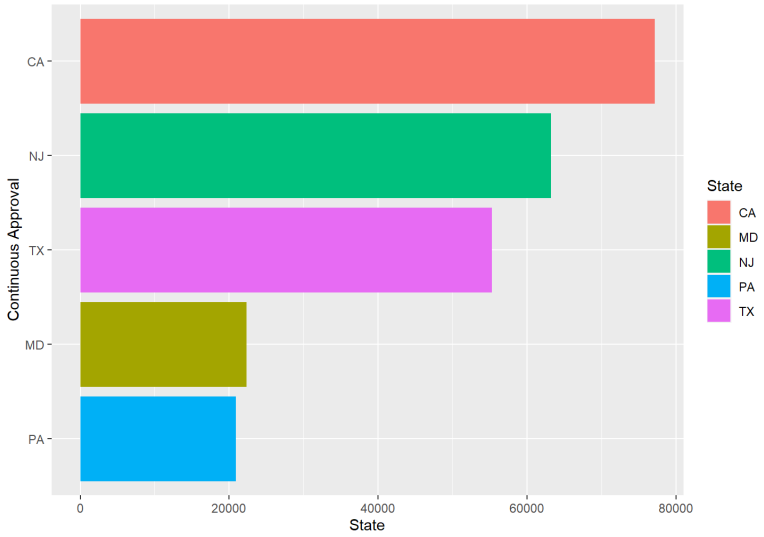
Visualizations

1)



The bar chart above shows the State from which most of the Initial Approval came from. Here as it can be seen Employers from Texas are the ones who approve most number of Initial H1Bs

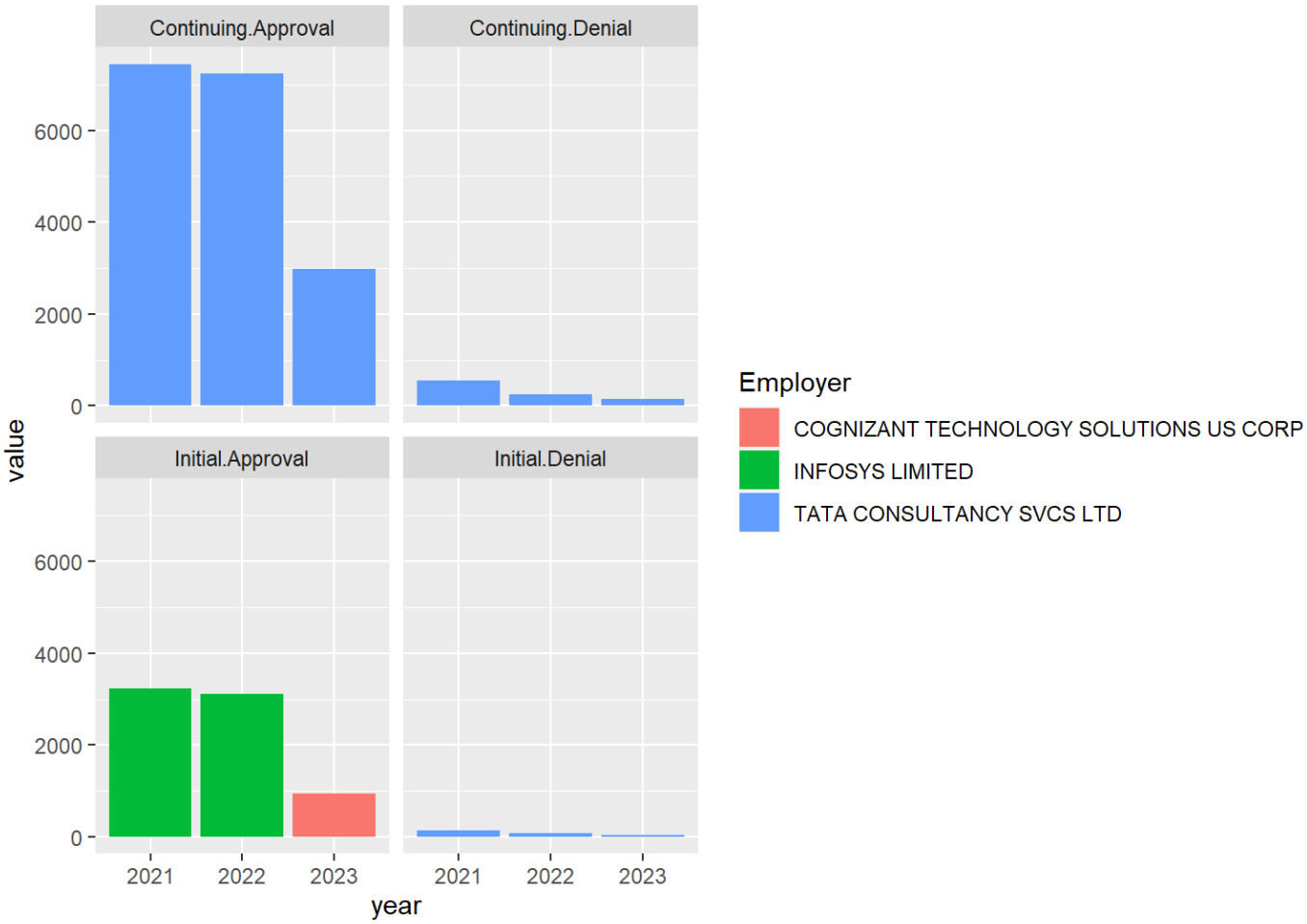
2)



The bar chart above shows the State from which most of the Continual Approval came from. Here as it can be seen Employers from California are the ones who approve most number of Initial H1Bs.

In both cases New Jersey seems to be a good options for H1B seekers as it ranks second for Initial and continual approval.

3)



The preceding bar chart showcases the leading H1B approvals/denial Employees across the year 2021,22,23. There's a common misconception that H1B-sponsoring companies predominantly hail from MANGA/FAANG, yet the data presents a different narrative. TCS ranks highest for continual approval, while Infosys leads in initial approvals with Cognizant emerging in 2023. Moreover, TCS exhibits a notable volume of denials, underscoring the sheer magnitude of applications they handle.

It is to be kept in mind that H1B approvals ultimately is a result of the US Government lottery system, the data provided above, helps us to understand the Employers that would readily accept the seeker and will have a higher chance of sponsoring one

Edited by **Anurag Akrathuveetil (https://northeastern.instructure.com/courses/170748/users/306209)** on Feb 16 at 5:20pm

Reply

Attach

Cancel Post Reply

Anagha Deshpande (https://northeastern.instructure.com/courses/170748/users/311259)

Feb 15, 2024

FLASH PAPER Anagha Source of the Data Set: The data set is from Centre for Disease control

FLASH PAPER

Anagha

1. Source of the Data Set: The data set is from Centre for Disease control and prevention. -- https://data.cdc.gov/ (https://data.cdc.gov/) . The Data contains the information of Provisional COVID-19 Deaths by Sex and Age in different states of America form the year 2004 to 2021. With the in deaths due to Covid-19 from the year 2019 to 2021, This dataset will be helpful to analyze how the death count differs from one state to another. In my visualization I have considered the death count for two states Georgia and California with aim of considering two states with fairly different migration rate.
   I wanted to visualize the death count over a given year "2021" in these two states and how it varies in different months. The intention of taking the death count for two different states is to check if that varies per month for two states situated at the west and east coast

2. The following is the structure of dataset :

Data As Of: Character variable representing the date when the data was last updated.
Start Date: Character variable representing the start date of the data period.
End Date: Character variable representing the end date of the data period.
Group: Character variable representing the grouping category.
Year: Numeric variable representing the year.

Month: Numeric variable representing the month.

State: Character variable representing the state.

Sex: Character variable representing the sex.

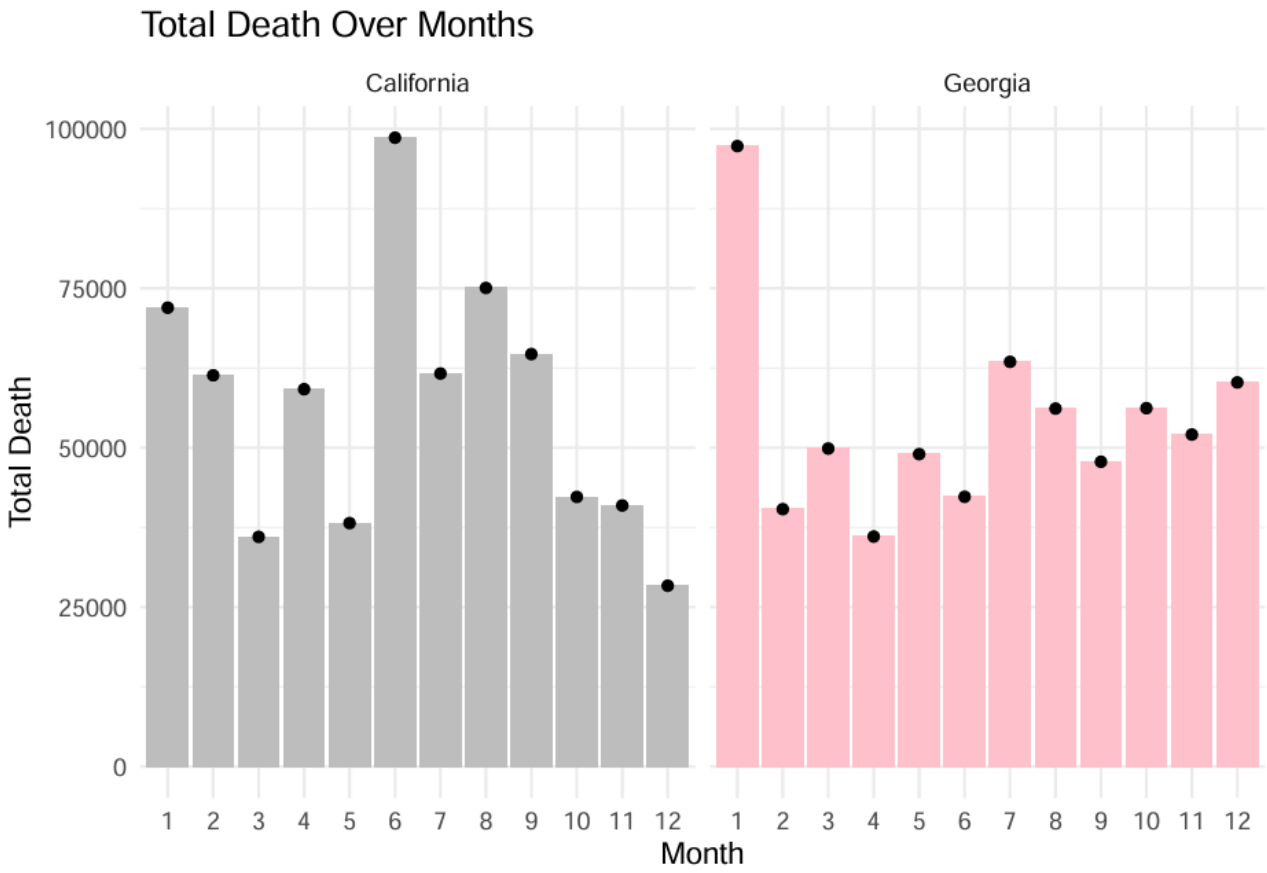
Age Group: Character variable representing the age group.
Year : Numeric Variable that represents the year.
States : Character variable that represent the State.

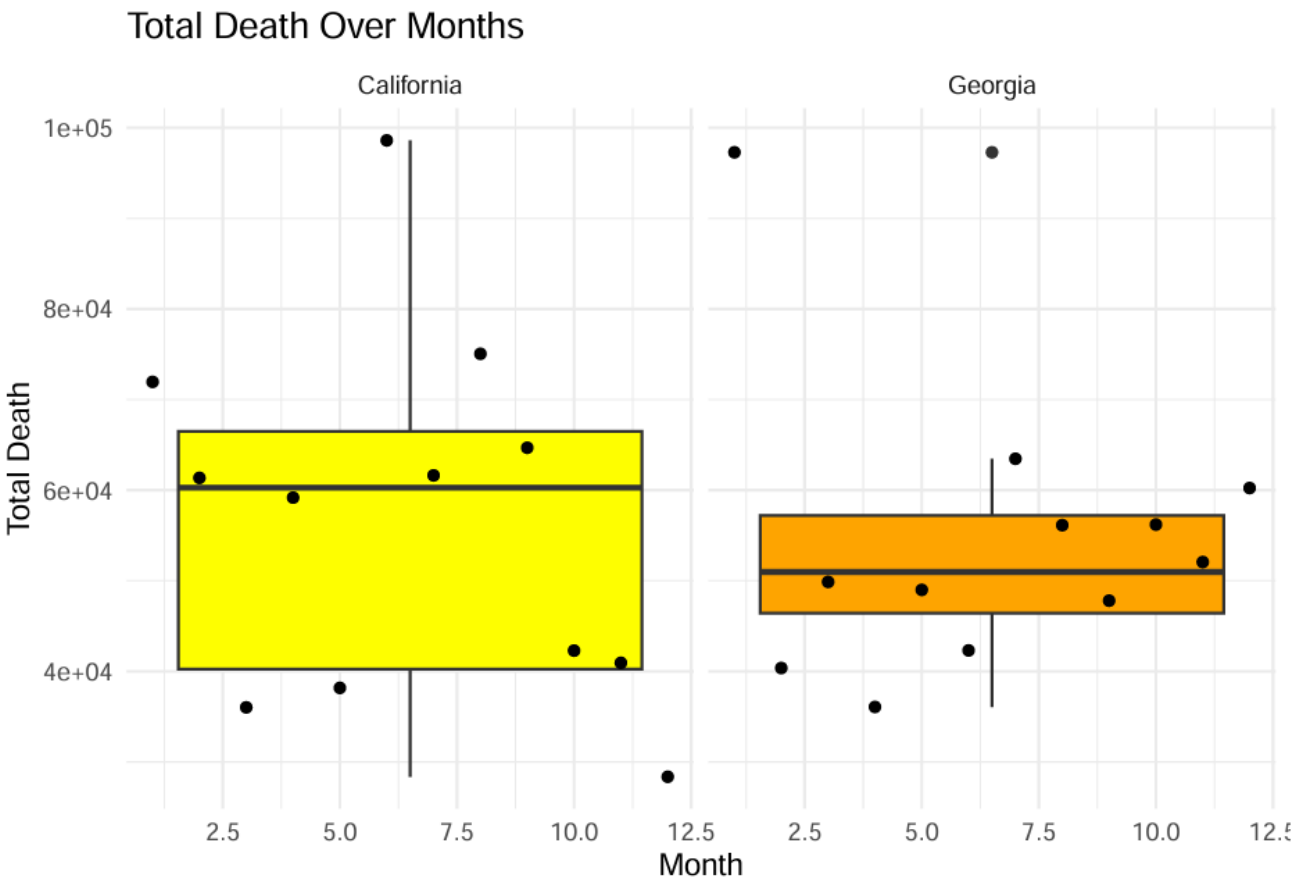
Total Death: Total number of Death for a given year in the country.

For the dataset above, I have considered the following variables of interest: Year, States, Total Deaths,
Data tidying by omitting the “NA” values and deleting the “Footnote” column from the dataset.
Used dplyr to clean the data and consider it only for the year “2021” for two states California and Georgia.
I have mutated the column Month by taking the sum of count of all the deaths for that column by summarizing the total deaths.

3.

3


Total Death Over Months
California
Georgia
Total Death
100000
75000
50000
25000
0
1 2 3 4 5 6 7 8 9 10 11 12
1 2 3 4 5 6 7 8 9 10 11 12
Month



The graph 1, in the above visualisation, gives the information about count of total death for every month in the year 2021 .
Georgia has had the highest death count in the month of January, and it has significantly reduced after the first month. However, for California, that is not the case. The highest death count is witnessed in the month of June for the year 2021. From this it is evident that the deaths are higher in high populated country like California with high migration rate.

Reply

Attach

Cancel Post Reply

Hiranmai Devarasetty (https://northeastern.instructure.com/courses/170748/users/312135)

Feb 15, 2024

FLASH PAPER 1. DATASET DES

FLASH PAPER

1. DATASET DESCRIPTION:

Source link: source (https://catalog.data.gov/dataset/conditions-contributing-to-deaths-involving-coronavirus-disease-2019-covid-19-by-age-group)

For Assignment 2, I chose Conditions contributing to COVID-19 deaths by state and Age Provisional 2020-23 dataset from the US Department of Health and Human Services. The Dataset compiles the information about health conditions and contributing causes mentioned in conjunction with deaths involving coronavirus 2019 (COVID-19) by age group and jurisdiction of occurrence. We know that COVID-19 transformed everyone's lives in every way, and I was intrigued by the number of COVID-19 deaths caused by different health conditions in various age groups. The Dataset also helps to get a clear picture of which health conditions worsened COVID-19 symptoms. My primary goal was to study the patterns and trends in COVID-19 deaths and understand how different variables contribute to a provision's overall COVID-19 death count.

2. VARIABLES AND STRUCTURE OF THE DATASET:

The Dataset had information regarding the COVID deaths and various variables such as State, Year, Month, Condition Group, Condition, ICD10 Codes, Age group, etc. After loading the Dataset, I observed a few outliers and missing values that needed attention. I performed data tidying,

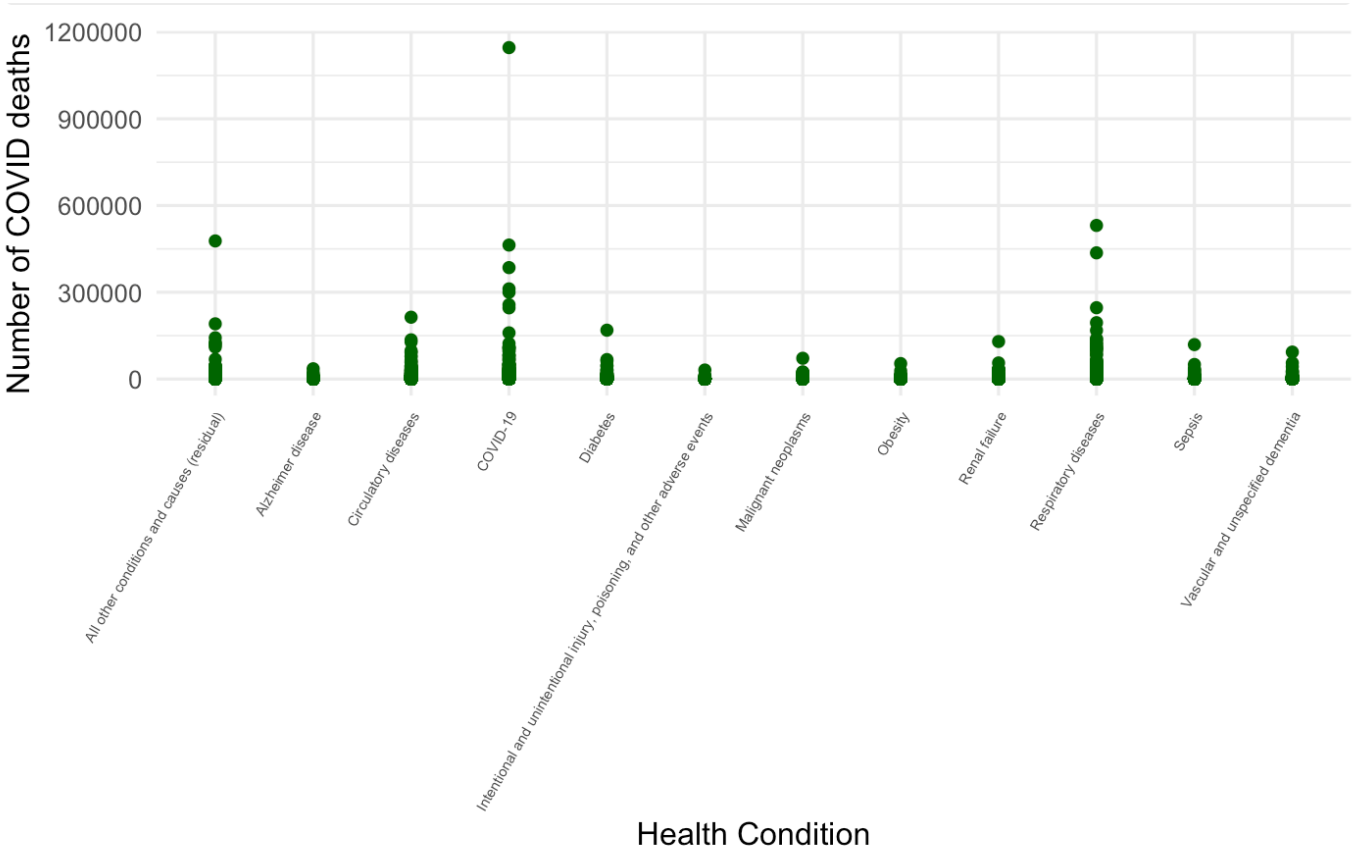
addressed the missing values, and tried removing the outliers to ensure the accuracy of the analysis. After analyzing the Dataset, a Few rows, such as Year and Month, imported NAs as their values, and the Flag row was empty. Hence, it was necessary to tidy the data by removing all these values as it impacts the accuracy of the visualization.

- ICD10 is the tenth version of the International Statistical Classification of Diseases and Related Health Problems. It is a medical classification given by the World Health Organization. The ICD10 Codes are utilized by healthcare providers and suppliers when submitting medical claims to Medicare. Hence, a different health condition in the dataset is given a different ICD10 code.
- The next exciting row in the dataset is the Condition Group, which speaks about the different health condition groups that affected the COVID death count.
- The Condition row has information on the different health conditions that affected the death count. This row contains the subgroups of the Condition Groups row in the Dataset.
- The state row shows the different provisions present in the United States of America.

3. VISUALISATION:

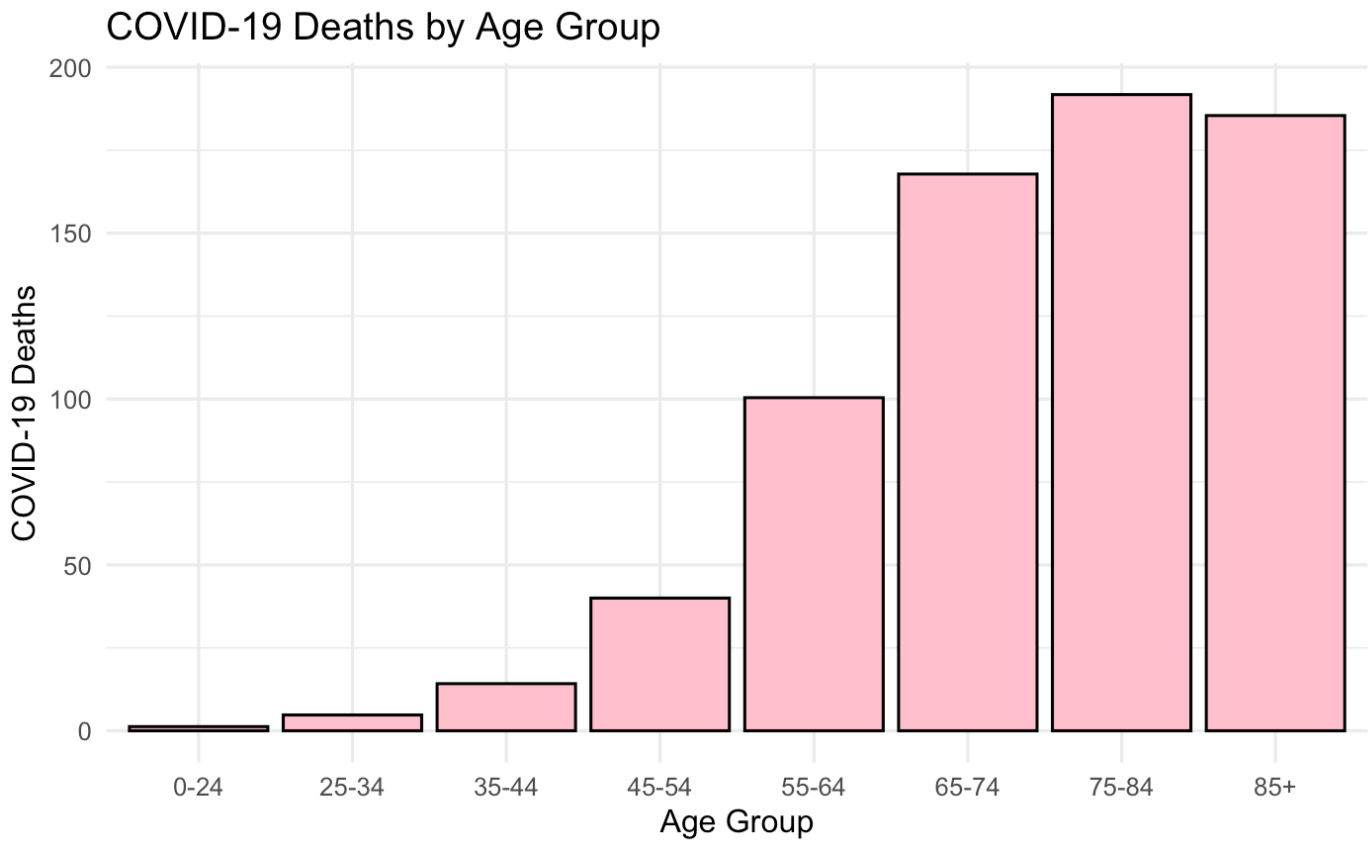
I have tried to focus mainly on the count of COVID deaths. Hence, I chose several COVID deaths as the basis on which all my visualizations revolve around.

1. One interesting aspect of my exploratory analysis is a plot graph comparing the Health Condition and the number of COVID deaths. This visualization showed the correlation between the Health Condition and the number of COVID-19 deaths.



From the above visualization, we can identify that most of the deaths were caused due to COVID-19 itself followed by the respiratory diseases. The graph shows that the different respiratory conditions fueled the coronavirus, leading to more deaths among the people in various provisions in the United States of America.

2. Another visualization compares COVID-19 Deaths with the age group. This helps gain insights into which age group was affected most by the COVID-19 and other health conditions.



The bar graph illustrates the age group in which most of COVID-19 deaths took place. The 75–84-year-old age group was affected the most by COVID-19 and the other health conditions caused by the Coronavirus, followed by the 85+ age group and 65-74 years. The graph clearly shows that the older people of age more than 65 years suffered a lot during the COVID period in the United States of America.

To summarize, the dataset helped gain insights into which health conditions triggered the Coronavirus and caused more deaths. Moreover, as the dataset contains data from the past three years, this also explains the different age groups affected by other provisions. Furthermore, visualizing the dataset concerning the State and working on the Coronavirus's side effects will help provide proper and early medical assistance to healthcare Providers to reduce the death count.

Reply

Attach

Cancel Post Reply

-

Gowreesh Gunupati (https://northeastern.instructure.com/courses/170748/users/245347)

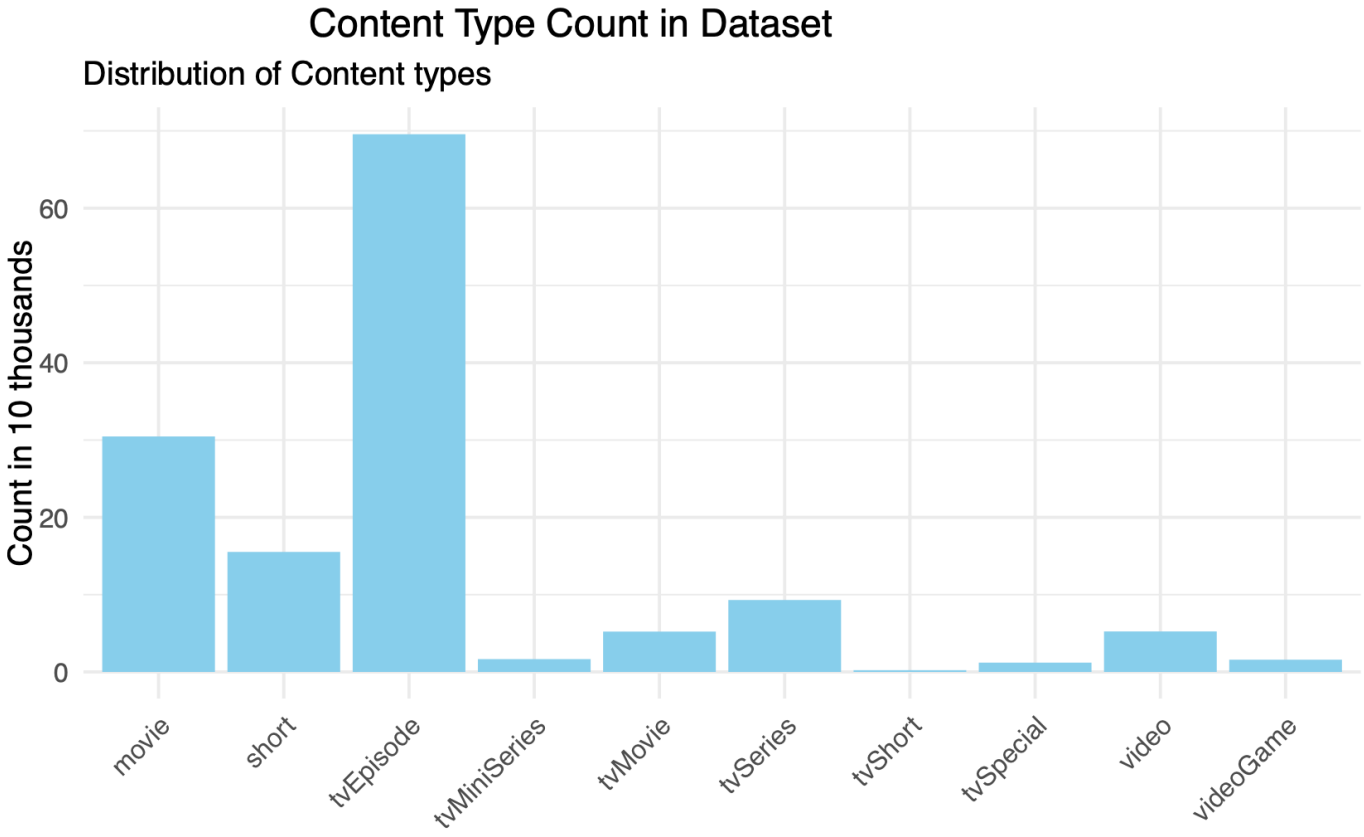
Feb 15, 2024

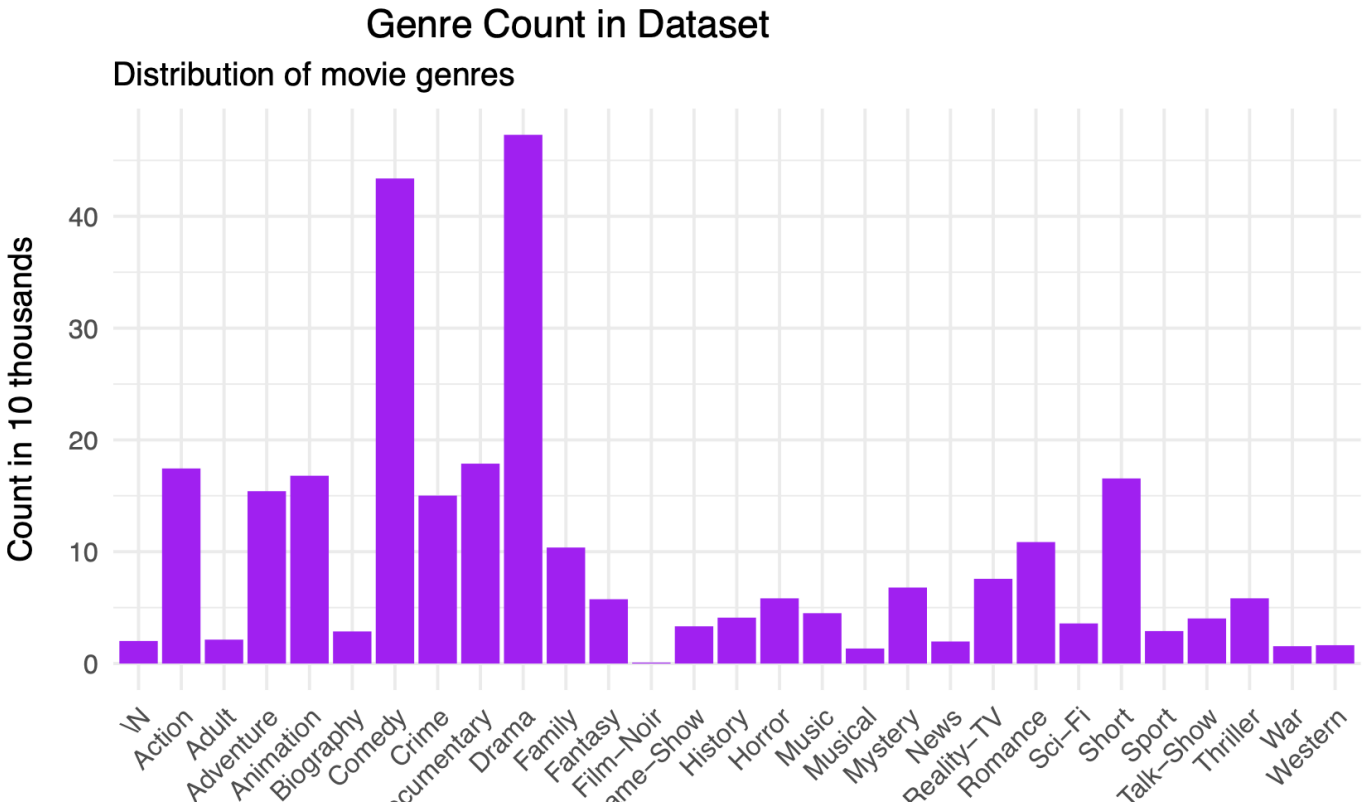
1. Dataset Description and Source: The dataset originates from IMDb, focusing on 'Title Basics' an

1. **Dataset Description and Source**: The dataset originates from IMDb, focusing on 'Title Basics' and 'Ratings'. It was selected for its comprehensive nature, allowing exploration of trends, genre popularity, and viewer ratings across entertainment content.

2. **Questions I wanted to explore:** I am someone who loves movies, and I have access to the IMDB dataset which contains movies from 1894 to 2024. I would like to create visualizations showing the distribution of genres, different content types, highest-voted and rated movies, how movies have been distributed over the years, and more.

3. **Structure and Preprocessing**: It includes details like title type, name, start/end years, and genres, alongside ratings. Preprocessing involved merging datasets, handling missing values, separating genres, and data transformation for analysis over different periods.


Content Type Count in Dataset
Distribution of Content types
Count in 10 thousands
0
20
40
60
movie
short
tvEpisode
tvMiniSeries
tvMovie
tvSeries
tvShort
tvSpecial
video
videoGame


Genre Count in Dataset
Distribution of movie genres
Count in 10 thousands
0
10
20
30
40
Action
Adult
Adventure
Animation
Biography
Comedy
Crime
Drama
Family
Fantasy
Film-Noir
History
Horror
Music
Musical
Mystery
News
Romance
Sci-Fi
Short
Sport
Thriller
War
Western

Reply

Attach

Cancel Post Reply

Kahan Dhaneshbhai Sheth (https://northeastern.instructure.com/courses/170748/users/287754)

Feb 15, 2024

Hello everyone, For the purpose of HW2, I selected the following problem statement. Problem Stat.

Hello everyone,

For the purpose of HW2, I selected the following problem statement.

Problem Statement: For this, I put myself in the shoes of the security chief for Los Angeles, we aim to use crime data to protect our community more effectively. This data helps us identify crime hotspots, optimize patrols, and allocate resources. It also aids in fostering collaborations with local businesses and residents for crime prevention. By leveraging this data, we improve security and the city's overall quality of life.

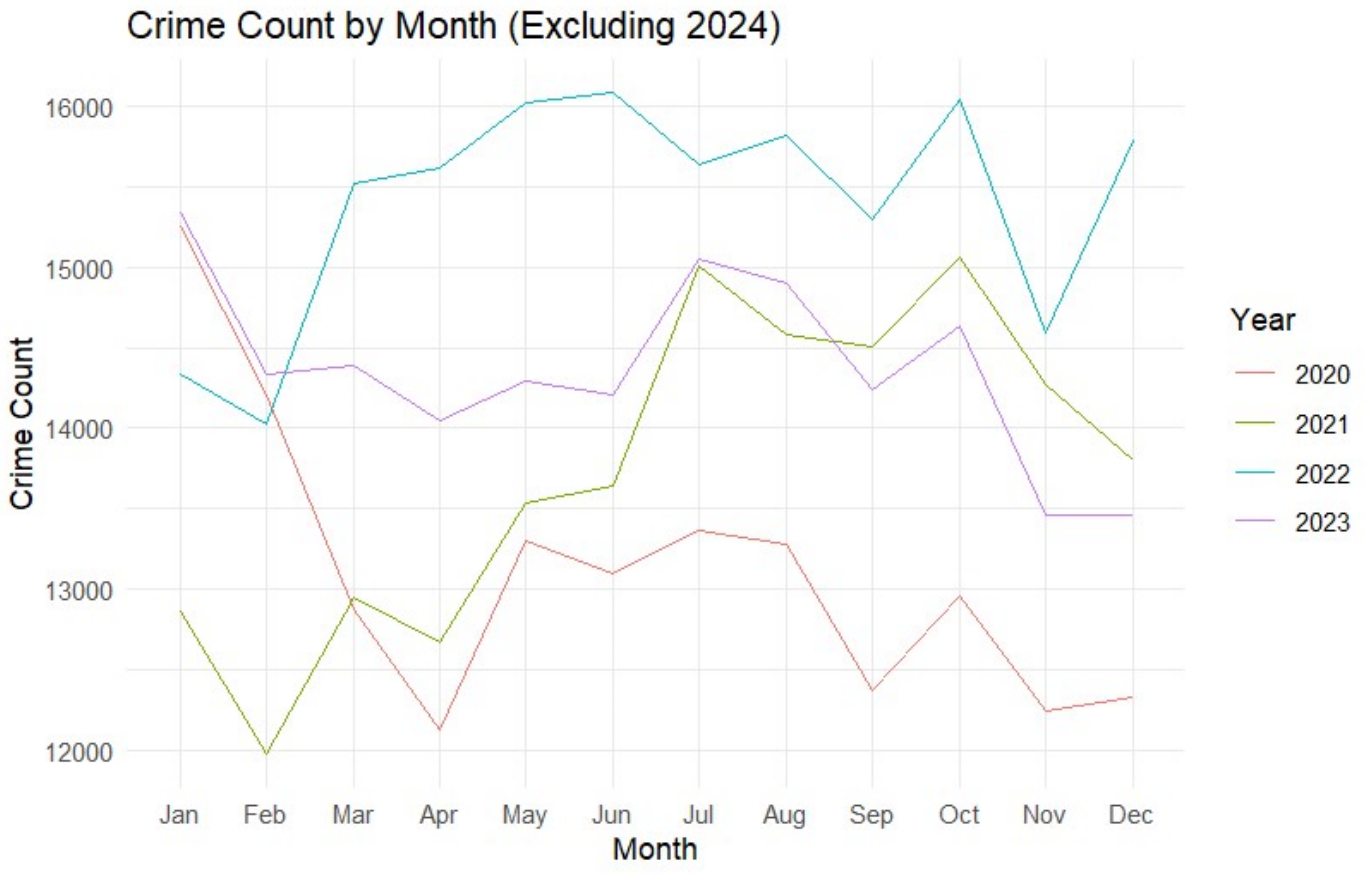
1. Dataset Description: The dataset used for this analysis is the "Crime Data from 2020 to Present" from the City of Los Angeles. This dataset is publicly available on the https://catalog.data.gov/dataset/crime-data-from-2020-to-present (https://catalog.data.gov/dataset/crime-data-from-2020-to-present) website. It contains records of crime incidents reported in the city, including details such as the date and time of the incident, the type of crime, location, and demographic information about the victims. I chose this dataset due to its relevance and importance in understanding crime patterns and trends in a major city like Los Angeles, additionally, as it is continuously updated, it provides the

most current and relevant insights into the city's crime landscape.

2. Dataset Structure and Preprocessing: The dataset consists of 883,987 observations and 28 variables. Some of the key variables include unique identifiers for each crime report, dates of occurrence and reporting, location information, type of crime, and demographic details of the victims. Before visualizing the data, preprocessing steps were necessary to handle missing values and ensure consistency. Columns with significant missing values were dropped, and values inconsistent with gender (other than 'M' or 'F') were converted to NA. Additionally, text-containing columns were preprocessed properly by performing essential text preprocessing steps like making them uppercased for uniformity.

3. Visualization and Exploration:

I present a line graph titled "Crime Count by Month (Excluding 2024)" which plots the crime counts for each month from 2020 to 2023. This graph is particularly interesting as it reveals distinct patterns and trends in the crime data.



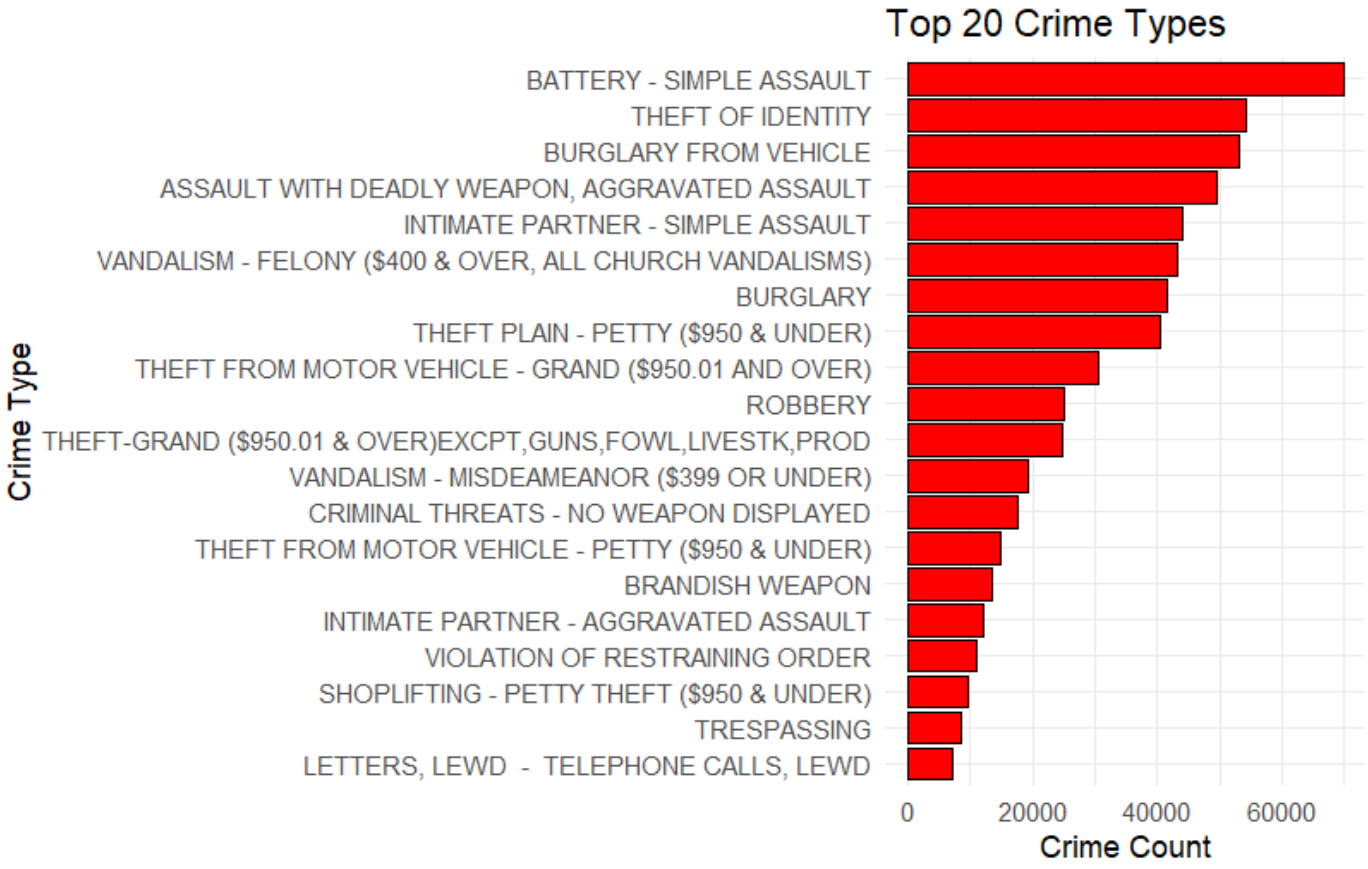
Observations from the graph include:

- **2020:** The crime count started at around 12,500 in January, dropped significantly around May (spring season), and peaked close to 14,000 in October (fall season).
- **2021:** The crime count began at nearly 10,000 in January, steadily increased until July (peaking at around 15,000), declined until September (end of summer), and rose again towards December (winter season).

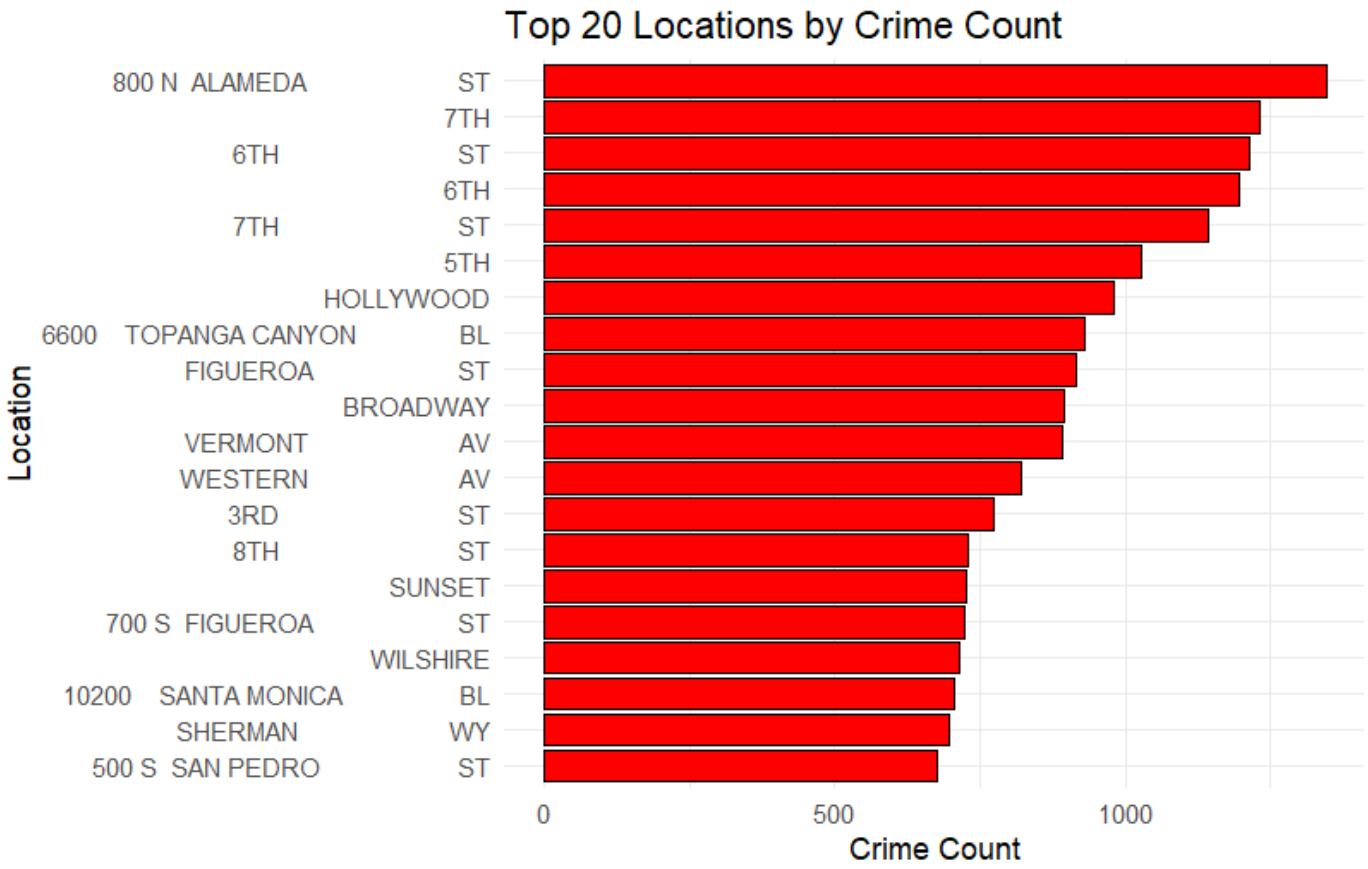
- **2022:** The year started with a high crime count of about 14,500 in January, sharply increased by April to around 16,000, and remained relatively stable for the rest of the year.
- **2023:** The crime count shows a relatively stable trend but with a downward trend towards the end of the year.

A notable observation across all four years is the peak in crime count during October, suggesting a potential seasonal or other factor that warrants further investigation. The data for 2024 was not included due to its limited volume, which would not provide reliable inferences.

Additionally, the bar graph titled “Top 20 Crime Types” provides a compelling visualization of the most frequent crime types in Los Angeles. The most common crime, as depicted, is “Battery - Simple Assault”, followed by “Theft of Identity” and “Burglary from Vehicle”. Other high-frequency crimes include “Assault with Deadly Weapon, Aggravated Assault” and “Intimate Partner – Simple Assault”.



The bar graph titled “Top 20 Locations by Crime Count” provides an insightful visualization of the locations with the highest crime counts in Los Angeles. The location with the highest crime count is “800 N ALAMEDA”, followed by “6TH ST”, “7TH ST”, “HOLLYWOOD BL”, and “6600 TOPANGA CANYON”.



When we combine this information with the earlier analysis on crime types, we can infer that areas like “800 N ALAMEDA” and “6TH ST” are not only high-crime areas but may also be hotspots for crimes such as “Battery - Simple Assault”, “Theft of Identity”, and “Burglary from Vehicle”. This information can be instrumental in guiding residents and visitors to exercise caution in these areas. Furthermore, this data can assist law enforcement in prioritizing resources and interventions for these high-crime areas, potentially reducing their prevalence and enhancing community safety.

This graph underscores the value of data visualization in making complex data more understandable and actionable. It provides valuable insights into the temporal patterns of crime in Los Angeles, informing strategies for effective resource allocation and crime prevention.

Thank you so much for reading!

Best Regards,
Kahan Sheth

Reply

Attach

Cancel Post Reply

Manasa Rao (https://northeastern.instructure.com/courses/170748/users/265275)

Feb 15, 2024

Electric Vehicle Population Data Dataset Description This dataset shows the Battery Electric Vehicl

Electric Vehicle Population Data

Dataset Description

This dataset shows the Battery Electric Vehicles (BEVs) and Plug-in Hybrid Electric Vehicles (PHEVs) that are currently registered through Washington State Department of Licensing (DOL).

Data Source: https://catalog.data.gov/dataset/electric-vehicle-population-data (https://catalog.data.gov/dataset/electric-vehicle-population-data)

I wanted to explore the market dynamics surrounding different electric vehicle makes and models, potentially to identify emerging trends or areas for innovation. Understanding the trends and patterns in electric vehicle adoption can provide valuable insights into the future of automotive technology and sustainability efforts. Moreover, as the world shifts towards renewable energy and low-emission transportation, having access to comprehensive data on electric vehicle populations can empower us to contribute to policy discussions, infrastructure planning, and environmental impact assessments.

This dataset was chosen to explore the specific aspects of the electric vehicle market. This makes it well-suited to exploring questions like number of electric vehicles by make, the distribution of electric vehicle types across different model years, the average electric range for different EV makes.

Structure of the dataset

The dataset is contained within a CSV file. It consists of rows and columns, with each row representing a specific electric vehicle entry and each column representing a different attribute/ characteristic of those vehicles.

Some variables of interest that is used in the visualization:

1)**Model Year:** The model year of the vehicle, determined by decoding the VIN(Vehicle Identification Number).

2)**Make:** The manufacturer of the vehicle, determined by decoding the VIN.

3)**Model:** The model of the vehicle, determined by decoding the VIN.

4)**Electric Vehicle type:** This distinguishes the vehicle as all electric or a plug-in hybrid.

5)**Electric Range:** Describes how far a vehicle can travel purely on its electric charge.

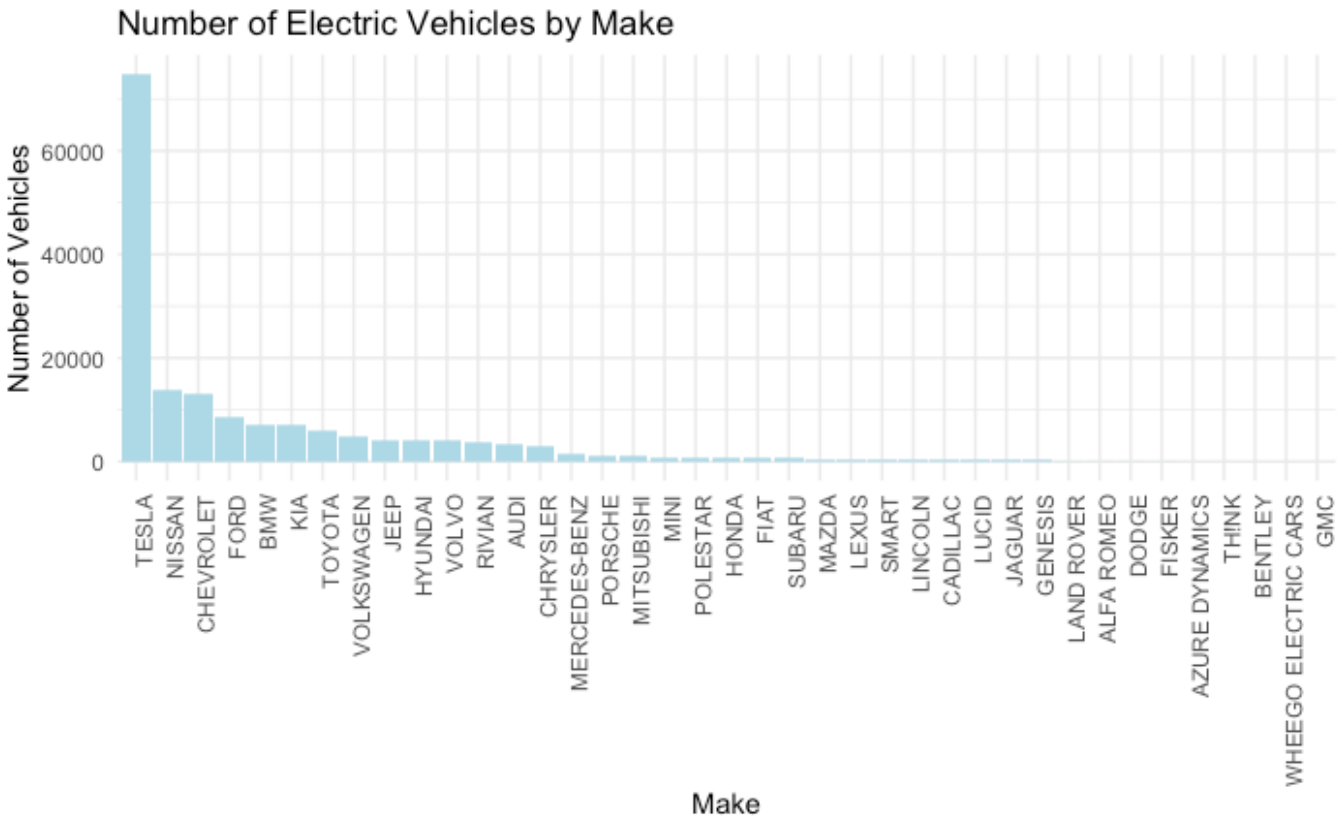
Data preprocessing needs and steps taken:

- Imported the dataset using read.csv() and loaded necessary libraries for data manipulation and visualization.
- **Handling missing values:** I have checked for missing values and identified that the columns Postal code, Legislative District and Census Tract contain missing data. Since these columns are not pertinent to my analysis, I've decided to ignore the missing values in these columns. This approach is reasonable given the analysis goals. However, it's important to ensure that key columns for my analysis i.e Make, Model Year, Electric Vehicle Type, and Electric Range do not contain missing values. If they do, we need to decide whether to impute these missing values or exclude the affected rows from our analysis.
- **Removing duplicate rows:** Duplicate rows can skew the analysis by artificially inflating the counts of certain vehicles. I've addressed this by removing duplicate entries, ensuring each row in the dataset represents a unique vehicle.

By performing these data cleaning steps, we can enhance the quality of our dataset, making it more suitable for analysis.

Visualizations

Number of Electric Vehicles by Make:



This plot serves as a crucial visualization for understanding the landscape of the electric vehicle (EV) market within the dataset. It provides valuable insights into which manufacturers are leading in terms of the number of electric vehicles registered.

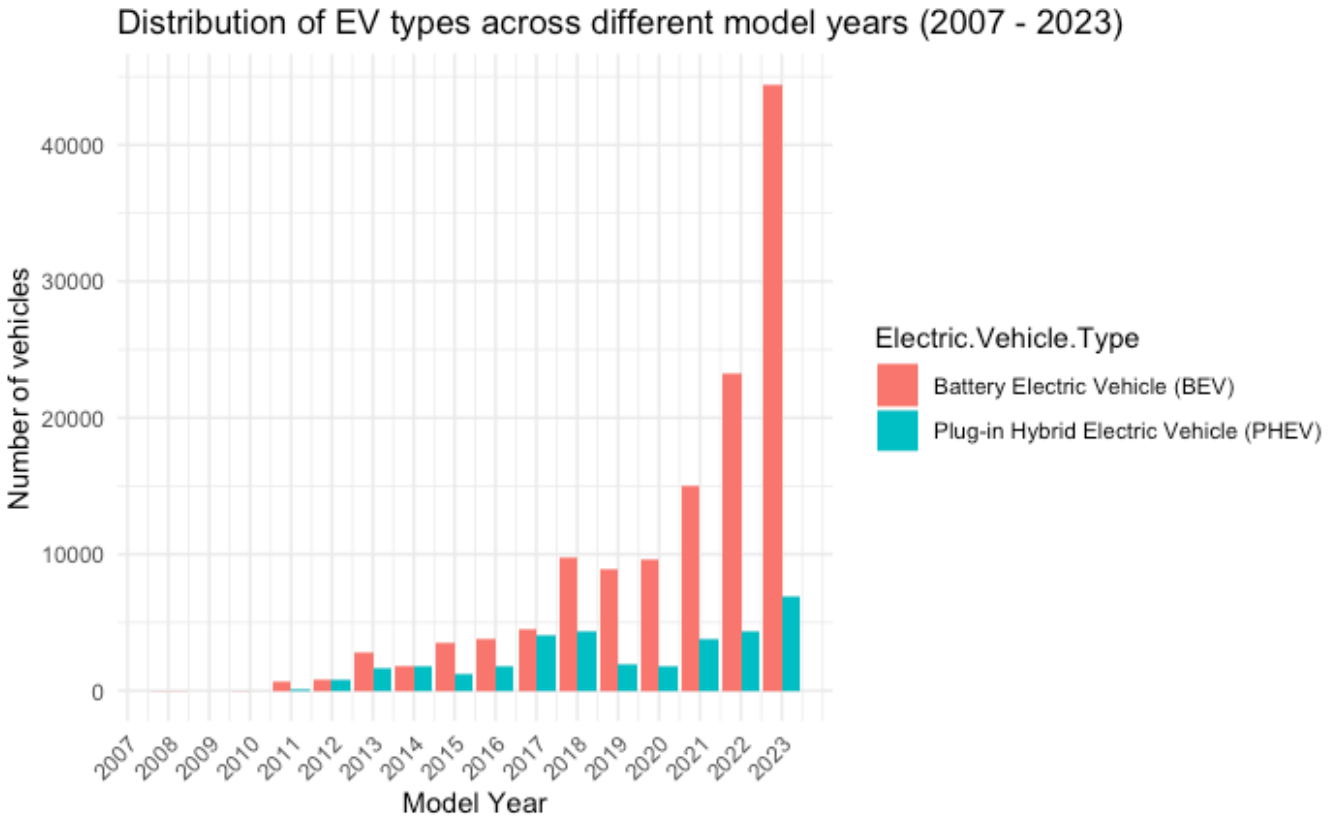
1)Tesla appears to be the most prominent make in terms of the number of EVs, with a significantly higher count than any other manufacturer. This suggests Tesla's dominant position in the EV market.

2)Following Tesla, there are other makes such as Nissan, Chevrolet, Ford etc, with a notable number of EVs, but their counts are significantly less compared to Tesla.

3)Many other makes have a relatively small number of EVs. These could represent manufacturers that are either new to the EV market or offer a smaller range of EV models

Tesla's leading position suggests that it is a market leader or a wider range of EV options provided compared to other manufacturers. The presence of multiple manufacturers indicates a competitive market with diverse options for consumers. However competition seems to be quite imbalanced. Therefore we can say that the EV market, based on this data, appears to be highly concentrated with a few key players dominating. For stakeholders such as manufacturers, investors this visualization can inform strategic decisions. Manufacturers can assess their market position and strategize on improving or maintaining their ranking. Investors might identify growth opportunities within the EV sector based on the popularity of certain makes.

Distribution of Electric Vehicle types across different Model years:



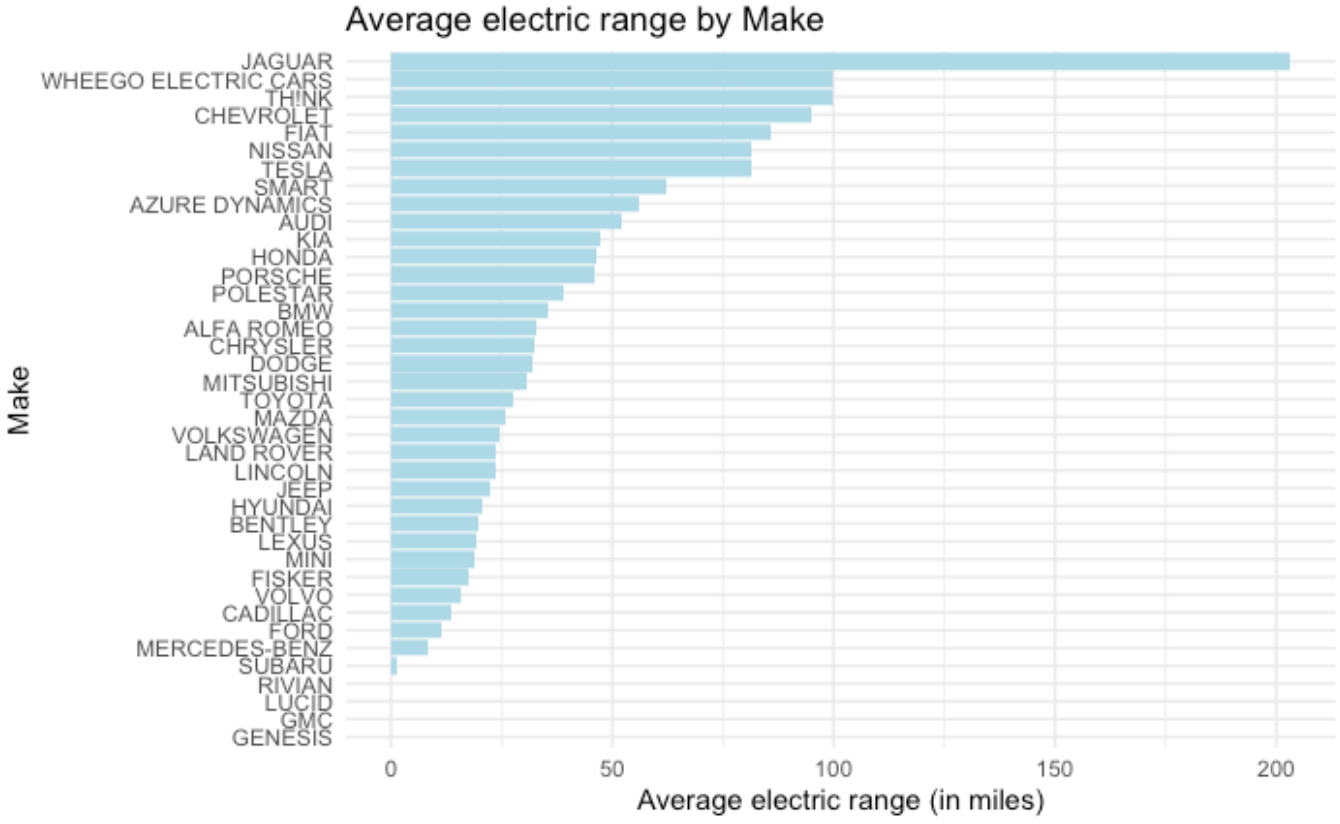
This plot illustrates the distribution of electric vehicle types (i.e., BEV, PHEV) across different model years. It highlights the evolution and trends in electric vehicle types over time, showcasing the growth or decline of each type.

1)There is a clear upward trend in the number of electric vehicles over time, with significant growth in recent years. This suggests increasing consumer adoption and possible improvements in EV technology and infrastructure.

2)The number of BEVs surpasses the number of PHEVs in almost every year, especially from 2018 onwards. This indicates a stronger market presence or consumer preference for BEVs.

Consumers and automotive industry appear to be shifting towards BEVs as the preferred type of electric vehicle. This might be due to a variety of factors. If the trend continues, BEVs may continue to grow in popularity. Specific years that show a noticeable increase in one type of EV over the other could mark technological milestones or the introduction of breakthrough models that significantly influenced consumer choices.

Average Electric Range for different EV Makes:



This plot is pivotal for understanding and comparing the performance capabilities of EVs from different manufacturers. A higher average electric range indicates that, on average, a manufacturer's EVs can travel farther without needing to recharge, which is a critical factor for many consumers concerned about range and the practicality of using an EV for their daily needs and longer trips.

1)There is a wide variability in the average electric range among different makes. Some

manufacturers offer vehicles with significantly higher average electric ranges than others.

2)The makes at the top of the chart, i.e Jaguar, Wheego, Th!nk has the longest average electric ranges.

This suggests that some manufacturers are potentially more focused on developing long range EVs than others. This might reflect their overall strategy and focus on the EV market. Brands with higher average electric ranges may appeal more to consumers who prioritize longer range for travel.

The dataset's geographical context could heavily influence the results. For example, certain makes might be more popular in specific region. Thus, the conclusions drawn from these visualizations are more relevant to the dataset's specific context and may not be universally applicable.

Edited by Manasa Rao (https://northeastern.instructure.com/courses/170748/users/265275) on Feb 15 at 7:48pm

Reply

Attach

Cancel Post Reply

-

Ruishan Shen (https://northeastern.instructure.com/courses/170748/users/248572)

Feb 15, 2024

Describe the dataset and where it comes from (making sure to cite the data source). Explain why y

Describe the dataset and where it comes from (making sure to cite the data source). Explain why you chose this dataset and what questions you wanted to explore in your visualization.

The dataset I chose is **"Students Performance in Exams"** from Kaggle, the website is https://www.kaggle.com/datasets/whenamancodes/students-performance-in-exams/data (https://www.kaggle.com/datasets/whenamancodes/students-performance-in-exams/data) . I picked this dataset because I want to explore the factors that affect the student's performance in the exams, such as parental level of education, lunch price, and the completion of test preparation course. Some questions I would like to explore are whether or not those aspects influence the student's overall

scores in math, reading, and writing and by how much.

Describe the structure of the dataset and the variables of interest. Describe any preprocessing needs (tidying, cleaning, transformation, etc.) and describe the steps you took to perform the preprocessing.

Data Description:

The dataset contains 1000 observations and 8 variables, are **gender**(the gender of students), **race/ ethnicity**(race groups from A to E), **parental level of education**, **lunch**(standard or free/reduced lunch), **test preparation course**(whether students completed or not), **math score**, **reading score**, and **writing score**.

Tidying the dataset:

The dataset's variables have long and redundant names that the tibble cannot show the last two columns from the left, so I used the "rename" to shorten the column names as education, prep, race, math, reading, and writing.

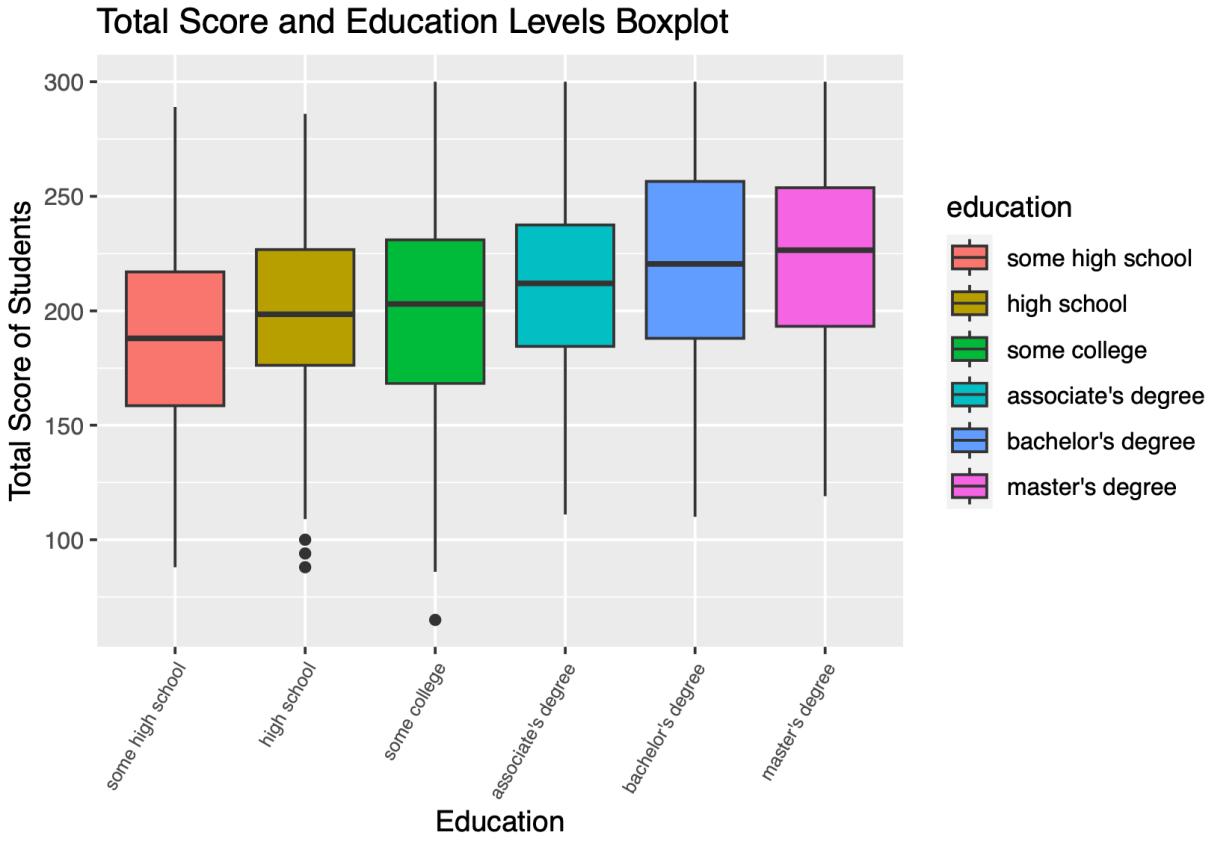
To see the student's performance more clearly, I use "mutate" to combine students' math, reading, and writing score as a total score.

Since we are more familiar with letter grades, it is clearer to classify the score levels using A, B, C, D, and F. I used "mutate" to set the range of total score greater than 180 as A, total greater than 160 as B, total greater than 140 as C, total greater and equal to 120 as D, and total less than 120 as F. It is easier to compare their performance by letter grades. Now there are two more variables: total and grade, which help with the observation.

There is no particular order in any variables, so in order to see how higher parental education level contributes to students' performance, I use "factor" to reorder the education levels from "some high school" to "master's degree".

Present at least 1 figure that is interesting to you and describe your observations and any key takeaways from the visualization and your exploration of the dataset.

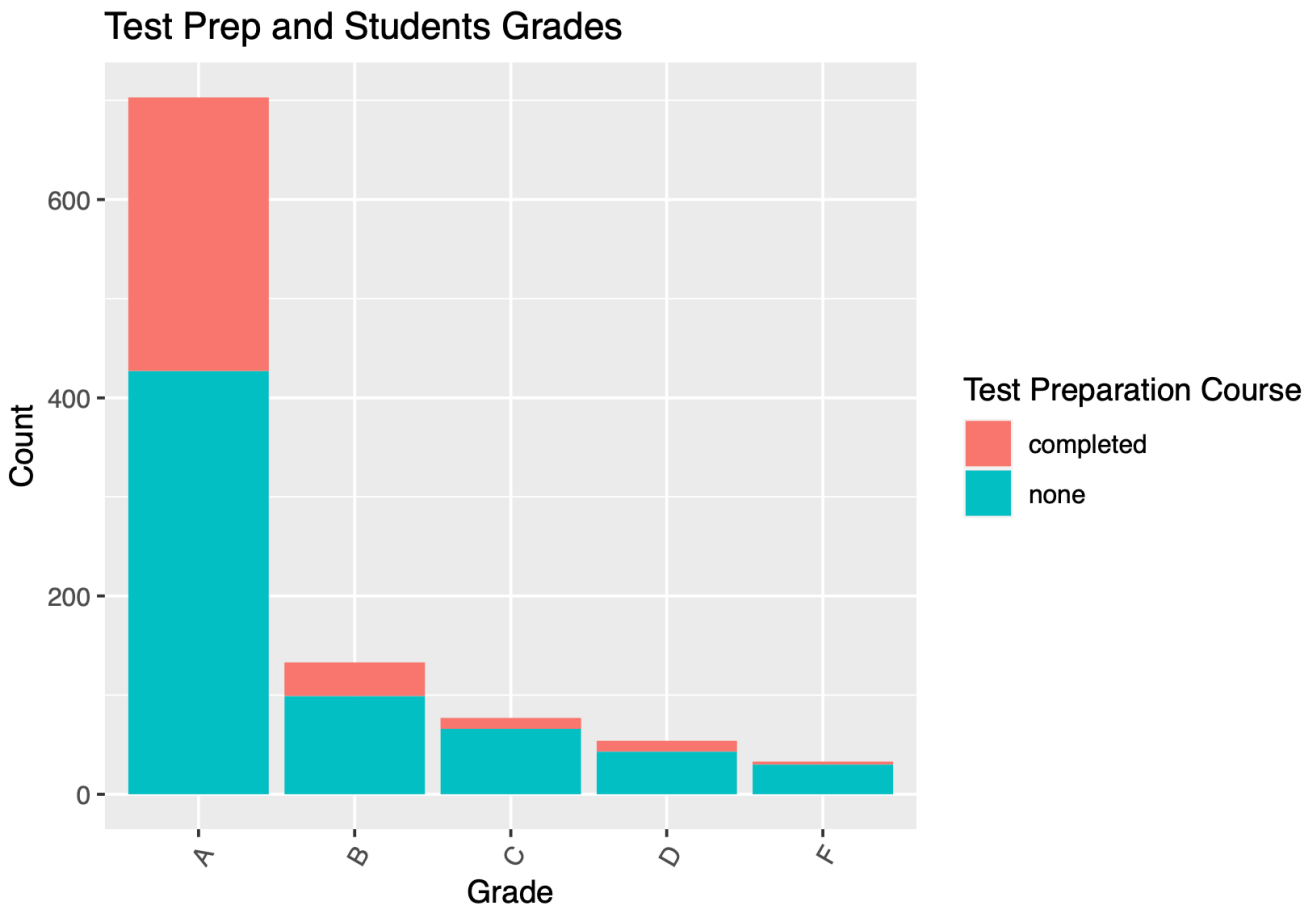
Total Score and Education Levels Boxplot:



I used a boxplot to show the relationship between students' average total scores and parental education levels. From the figure, we can see the average of total scores is increasing as the parental level gets higher. The higher the education levels of parents, the higher the average total scores of students, probably because parents with higher education levels tend to focus more on the education of their children.

There are also obvious outliers and a greater range of scores for educational levels that are lower than "associate's degree".

Test Prep and Students Grades:



I use a bar chart to show the relationship between students' grades and their test prep completion. From the observation, the higher the grade, the greater the portion of students completed their test prep course. It seems that the test prep can help students get better scores.

Reply

Attach

Cancel Post Reply

-

Gargi Girish Umrajkar (https://northeastern.instructure.com/courses/170748/users/230898)

Feb 15, 2024
DATASET DESCRIPTION I chose the IMDb dataset from Kaggle - https://www.kaggle.com/dataset

DATASET DESCRIPTION

I chose the IMDb dataset from Kaggle - **https://www.kaggle.com/datasets/bharatnatrayn/movies-dataset-for-feature-extracion-prediction (https://www.kaggle.com/datasets/bharatnatrayn/movies-dataset-for-feature-extracion-prediction)** . I chose this dataset for its comprehensive and multi-dimensional nature, knowing it would present a substantial learning curve in data cleaning and feature extraction. As a beginner in data science, I was looking for a challenge that would give me hands-on experience with real-world data that is often unstructured and requires a great deal of preprocessing.

Some specific questions I wanted to explore through visualization and analysis include:

1. How are ratings distributed across different types of entertainment, such as movies versus TV shows and across various genres?
2. What are the total gross earnings attributed to each director within the dataset, and how do these earnings correlate with the director's filmography?
3. What is the distribution of ratings with respect to the number of votes received, and is there a noticeable trend or correlation between a higher number of votes and higher ratings?

DATASET STRUCTURE AND PREPROCESSING

The dataset itself contains a wealth of information that spans various aspects of movies and TV shows on Netflix. With columns like MOVIES, YEAR, GENRE, RATING, ONE-LINE, STARS, VOTES, RunTime, and Gross, it encapsulates details ranging from quantitative metrics, such as ratings and box office earnings, to qualitative descriptions like genre and plot summaries. This diversity allows for multifaceted analysis and the opportunity to practice different types of data manipulation and visualization techniques.

The parameters in the dataset are as follows:

MOVIES: The title of the movie or series.
YEAR: The release year or the range of years the movie or series was active.
GENRE: The genre(s) associated with each title, which could include multiple genres for a single title.
RATING: The average rating given to the title, on a scale typically from 1 to 10.
ONE-LINE: A brief description or tagline for the title.
STARS: Initially contained information about the director(s) and main stars, including the prefix "Director:" for directors and "Stars:" for leading actors/actresses.
VOTES: The number of votes the title received.
RunTime: The duration of runtime of the movie or 1 episode of the series.
Gross: The gross earnings or box office collection.

The preprocessing and cleaning steps I undertook for this dataset were essential to ensure its usability for analysis and visualization. Here's a summary of what I did:

1. I started by loading the dataset from a CSV file into R, utilizing the `read.csv` function, which is a common starting point in data analysis.
2. In the YEAR column, I removed parentheses to clean the data while retaining its original form, which would allow me to easily create new variables later on.
3. To better categorize the data, I introduced a new column named “Entertainment Type.” This helped in classifying each title as either a “Movie,” “On Air Series,” “Off Air Series,” or “Unknown.” This classification was based on the information provided in the YEAR column.
4. I trimmed all leading and trailing whitespace from character columns. This step is crucial to avoid errors in text analysis and when performing operations that rely on string matching.
5. I also removed newline characters from all character columns to maintain a standardized text format across all data entries.
6. To handle missing values in the RATING and Gross columns, which could disrupt statistical calculations, I filled them with the mean of their respective columns. This is a common technique for imputing missing numeric data.
7. In the Gross column, I stripped away the currency symbols and abbreviations, converting all figures to a consistent numeric format indicating millions of dollars. I also renamed the column to “Gross in $-M” for clarity.
8. From the STARS column, I extracted director information and created a separate “Director” column. I removed the “Director:” prefix in the process and cleaned out the director information from the STARS column, leaving it to contain only actor names.
9. Finally, I cleaned up the STARS column by removing the “Stars:” prefix, leaving just the names of the actors and actresses, which makes the data ready for analysis involving cast members.

Through these meticulous steps, I ensured the dataset was structured and standardized, paving the way for accurate and insightful data analysis. These steps not only cleaned the data but also enriched it by creating new variables that could provide additional insights during the analysis phase.

VISUALIZATION

The horizontal bar plot depicts the total gross earnings of the top 20 directors, ranked by their financial success in the film industry. Each bar represents the gross earnings in millions of dollars, offering a straightforward comparison and highlighting the commercial achievements of these directors. This visualization is valuable for identifying key players in the film industry, understanding market trends, and examining the correlation between financial success and other factors like film ratings and genres.

CONCLUSION

In conclusion, the visual analysis of the IMDb dataset underscores the commercial success of certain directors and provides insights into the varied preferences of viewers. This investigation sets the stage for deeper explorations into the dynamics between a film's attributes—such as genre, ratings, and star power—and its box office performance, enriching our comprehension of the entertainment industry and consumer trends

Edited by **Gargi Girish Umrajkar (https://northeastern.instructure.com/courses/170748/users/230898)** on Feb 15 at 8:32pm

Reply

Attach

Cancel Post Reply

- 1
- 2