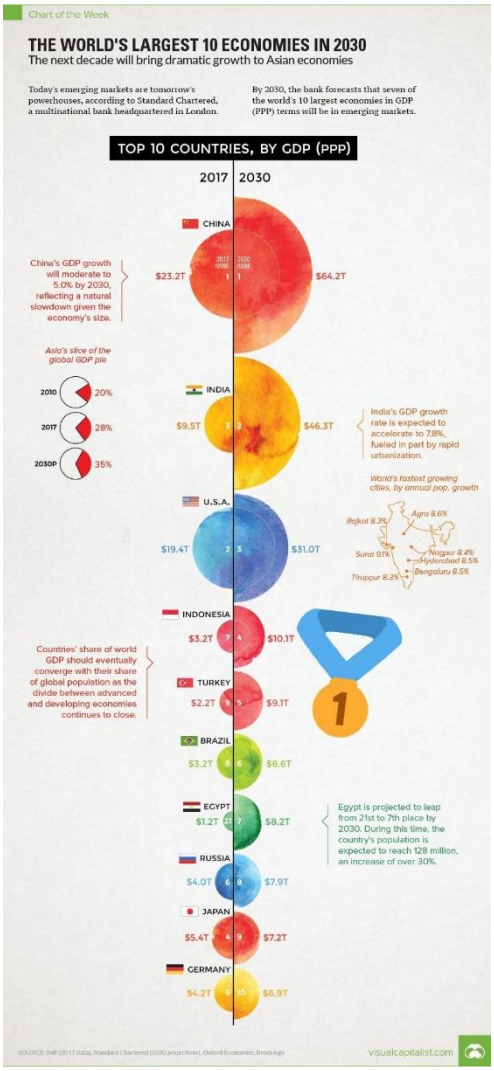
STAT 515 – MID PROJECT

BAD GRAPH OR TABLE REDESIGN PROJECT

1 INTRODUCTION

Misrepresentation of data, intentionally or unintentionally, is a prevalent problem in the current world where data is used to achieve nearly anything from finding the name of the song which you don't remember to training an Artificial intelligence such as Dall-E to produce realistic images just form a sentence. Tackling misrepresentation of data is important because it can be used to mislead people into beliefs that aren't true, might not be able to prove a point that the data is supposed to, the graph might miss key elements of data or have excessive decorations which makes it difficult to read etc. During this project, we have corrected three bad visualizations that misrepresent the data to a varying degree using R, an open-source programming language, and various libraries included with it.

2 REDESIGNS



2.1 Redesign 1

The First bad graph used was sourced from Visual Capitalist, an online publishing website. The graph is part of an article titled "The World's Largest 10 Economies in 2030", where, as the title suggests, the economic projections of 10 largest countries given by the Standard Chartered for the year 2020 which were discussed in the CEOWORLD magazine were quoted. The visualization here is trying to portray the countries which will have highest Gross Domestic Product (GDP) in trillions of dollars, calculated using purchasing power parity (PPP) measures.

2.1.1 Problems with the visualization

The primary issue with the graph is that circles are used to represent the data, which is makes it very difficult to observe the difference between the current and predicted value. The changes between different countries also diminish due to the circle's effect. The gradient used, while making the graph interesting to look at, worsens the problem of being able to differentiate between the two comparisons being made. The 5^{th} to 9^{th} ranking countries show almost no difference. The secondary issue is that the data was over-rounded to make the graph look attractive loosing a lot of data and changing some of it in the process.

2.1.2 Redesign goals:

The purpose of this redesign is to bring back the clarity lost due to the use of circles by plotting a bar graph instead and use the original data from the sources mentioned by the author of the graph and compare the results.

2.1.3 Redesign process:

Data sourcing:

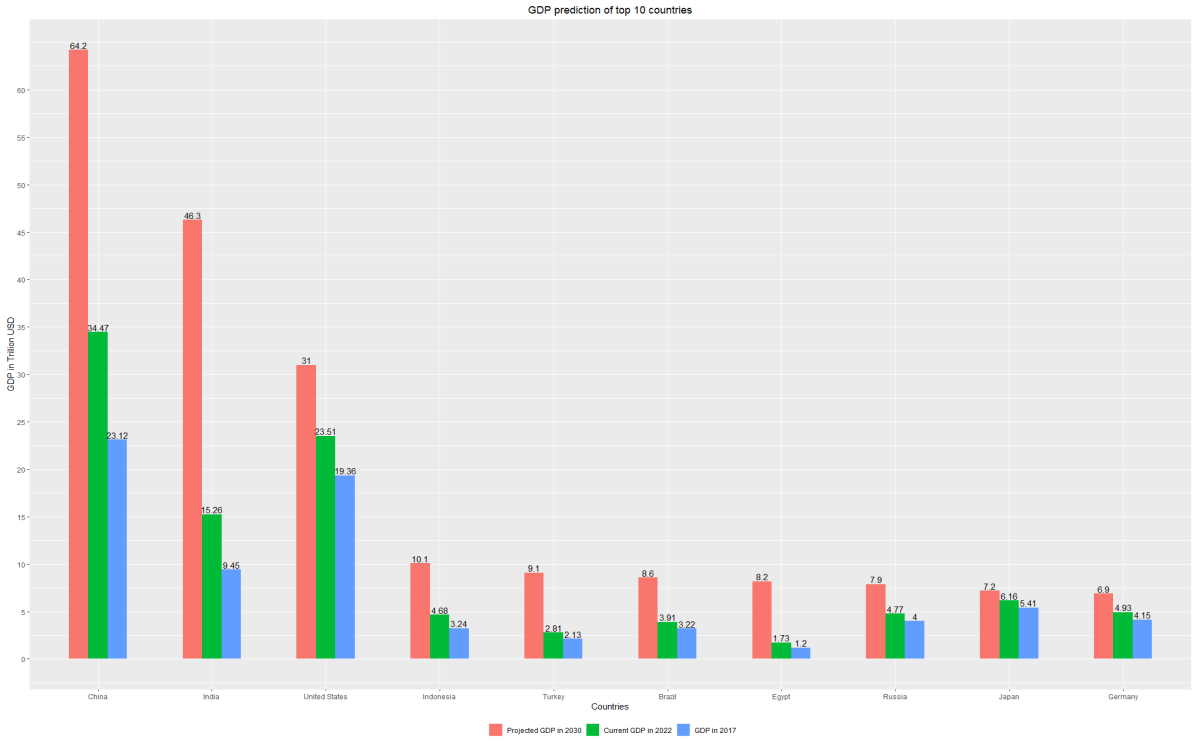
The data contains of two parts, one is the current GDP based on PPP from International Monetary Fund website and the other part of the data which is the projections is sourced from the article by CEOWORLD magazine titled "Standard Chartered: By 2030, These 10 Economies Will Be The World's Largest" by Sophie Ireland.

Data preparation:

The data from the sources was imported into .xlsx and the rows containing the GDP by PPP for the projected countries were selected for the years 2017 and 2022 and were merged into one single data frame for the visualizations to be produced.

Graph design:

The data for the countries is categorical and the GDP in trillions of USD was of a continuous scale. Considering these parameters, a bar graph, even though the most boring graph, seems to be the ideal choice because of the simplicity of and accuracy of representation of linear change in the consecutive categories. We designed the following grouped bar chart using function like ggplot, geom_col etc., from tidyverse package.



2.1.4 Challenge

The data source mentioned was very vague so finding and matching the dataset took extensive research of the websites of IMF and many other commerce websites. The data collected was unclean because it had data of 8687 rows and 53 columns of which only 10 rows and 2 columns were necessary, and the numbers noted had commas typed into them as strings instead of being of numeric type, so they had to be cleaned using various libraries and the final data was made into a new excel sheet to be imported into RStudio for further processing. The processing and importing were even harder because the dataset was 5 years older and the format of excel (.xls) used wasn't being fully supported by R now. The data used for the actual visualization being rounded or slightly changed made it more difficult to match the datasets and be confident that the datasets are indeed the same.

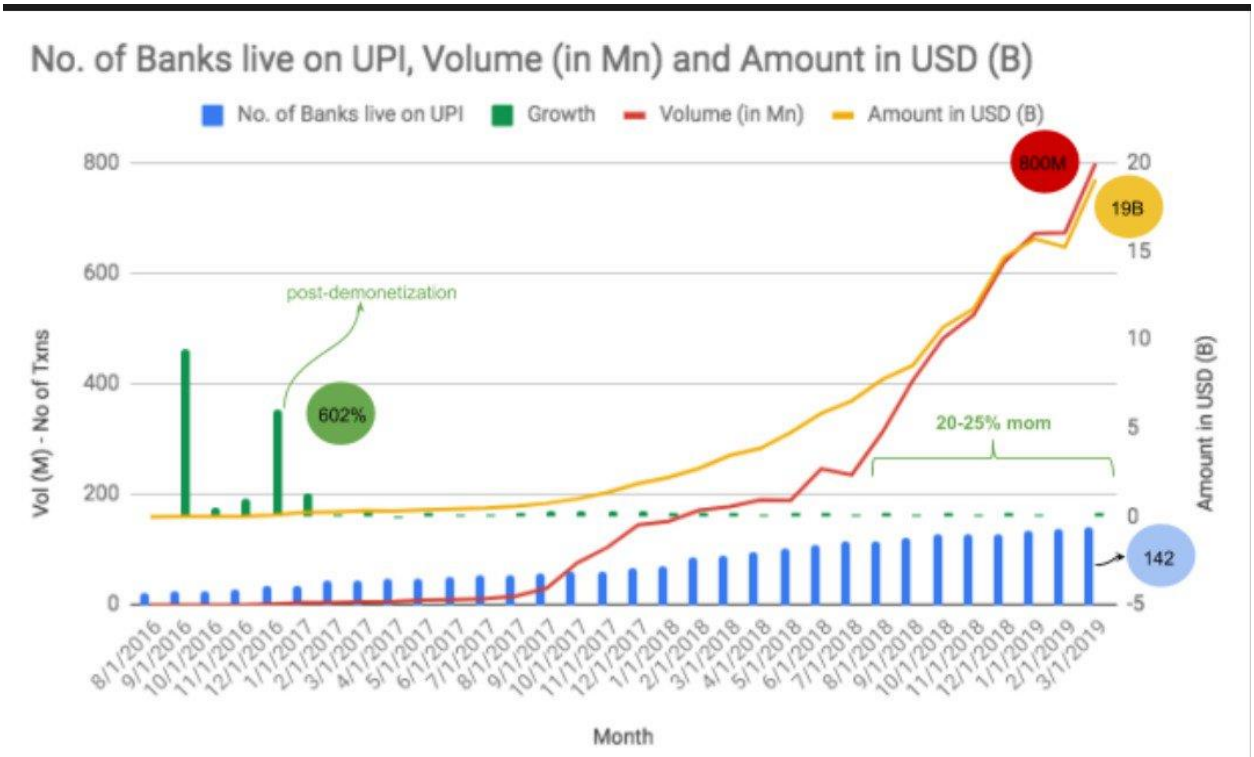
2.2 Redesign 2

This graph was sourced from a twitter post by @MohapatraHemant where the success of Unified Payment Interface (UPI) for touching 19 billion in USD. The visualization couldn't convey this because of a lot of reasons which are discussed further into the paper.

2.2.1 Problems with the visualization

The primary problem with the visualization is that it has a lot to show but ran out of space. All the lines, bars are crammed onto one plane and even though everything has its own scale, an

attempt was made to mix all the scales together, which made the volume that is in a few 100 million to cross the 19 billion line of the value of transactions. The percentage growth was plotted within the same scale of already confusing millions and billions. The number of banks in 100's was also implemented into the same graph but was somehow shown legibly which makes it misleading. The scale of growth and volume of transaction starts at over 180 million and has its maximum value matching 800 million, which is very misleading.



2.2.2 Redesign goals

The redesign of this graph was done with a primary goal of showing as much data as possible with actual non-manipulated scales and see how they compare with each other while looking for any trends within the data available.

2.2.3 Redesign process

Data sourcing:

The data for this visualization was sourced from National Payment Corporation of India's (NPCI) website. Because the visualization was old it contained the data only until the year 2019, but we had the data until September 2022, so all the data was used to provide latest context and see how the trends have panned out.

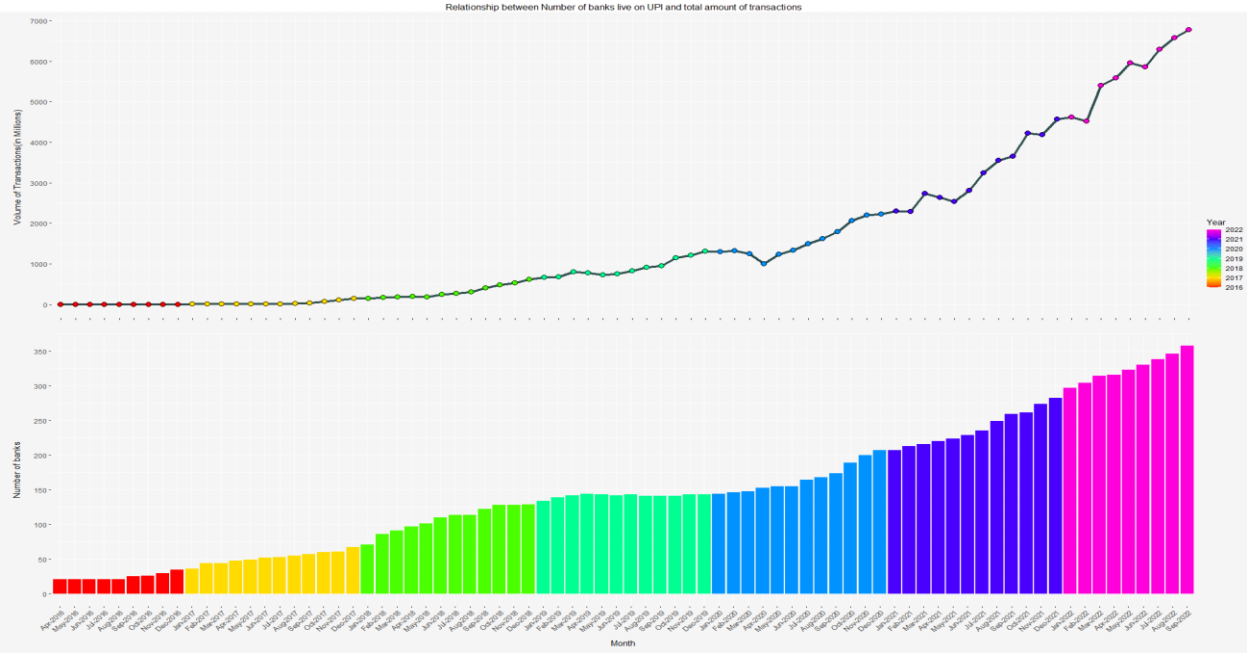
Data preparation:

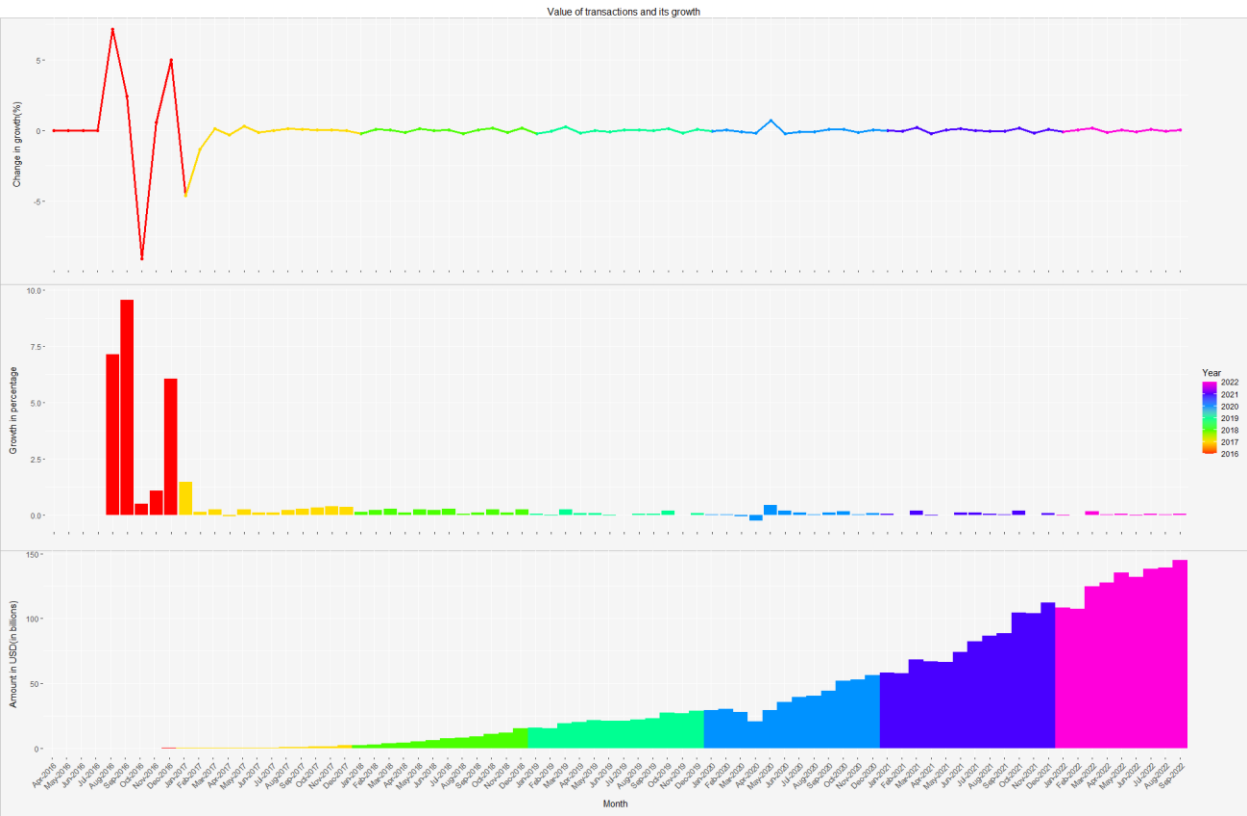
The data obtained was simple with 4 columns of data that need to be plotted. The percentage growth and the change in growth had to be calculated based on the given data of value of transactions. The data also had to be converted into units used in the visualization and had to be reduced by a degree allover to fit the data of all the new years.

Graph design:

Graph 1 of the redesign contains the number of banks that are live on UPI and the volume of transactions versus the months of years 2016 to 2022. The data was readily available and the data being categorical made bar graph using geom_bar seemed like the ideal choice. The data about the transactions had a pattern and the data points were not too many so a scatter plot using geom_point overlaying a line using a geom_line was used.

Graph 2 of the redesign contains three directly related variables that is the amount of transactions in billions of USD, the growth in percentage of amount of transactions, the change in growth that wasn't the part of the actual graph was introduced to show how the growth over time has scaled. A simple bar graph using geom_col was constructed for both growth in percentage and value of amount in USD because the data trends needed to be identified and the difference between each point of reference had to b shown with as much precision as possible.





2.2.4 Challenge

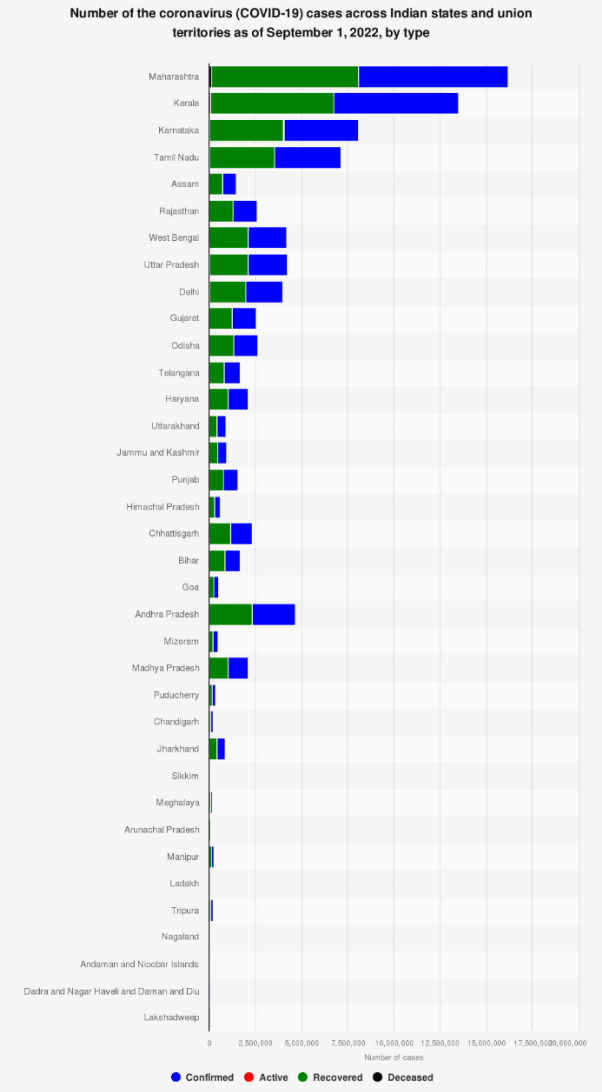
The source of the visualization fails to mention the source of the data so finding the was a very difficult task. Then based on intuition we tried to find all the datasets related to UPI and landed in the NPCI website which was the organization that owns UPI and managed to find a dataset that had similar entries. But the data set hosted in the website had taken great measures to prevent the data from getting out so we had to convert the HTML code of the table within the website to excel. The data obtained was in INR because that is the currency they deal with, so the data has to be converted to USD based on the specific year's average separately because the original graph had the value in USD.

2.3 Redesign 3

This project's graph was pulled from the Statista website. The bar graph displays the overall number of covid cases, including those who have passed away, been discharged, and are still active. The data includes all cases that have been reported and are organized according to Indian states and union territories.

2.3.1 Problems with the visualization

The stacked bar graph makes it difficult to identify the active and deceased cases because the magnitude of all the categories is very different, and the sorting for the states isn't oriented to produce and useful effects, these issues prevent a better visualization and explanation of the data and the situation.



2.2.2 Redesign goals

The redesign was oriented in such a way that the lower order of magnitude data would be shown properly because the data was significant enough to be ignored.

2.2.3 Redesign process

Data sourcing:

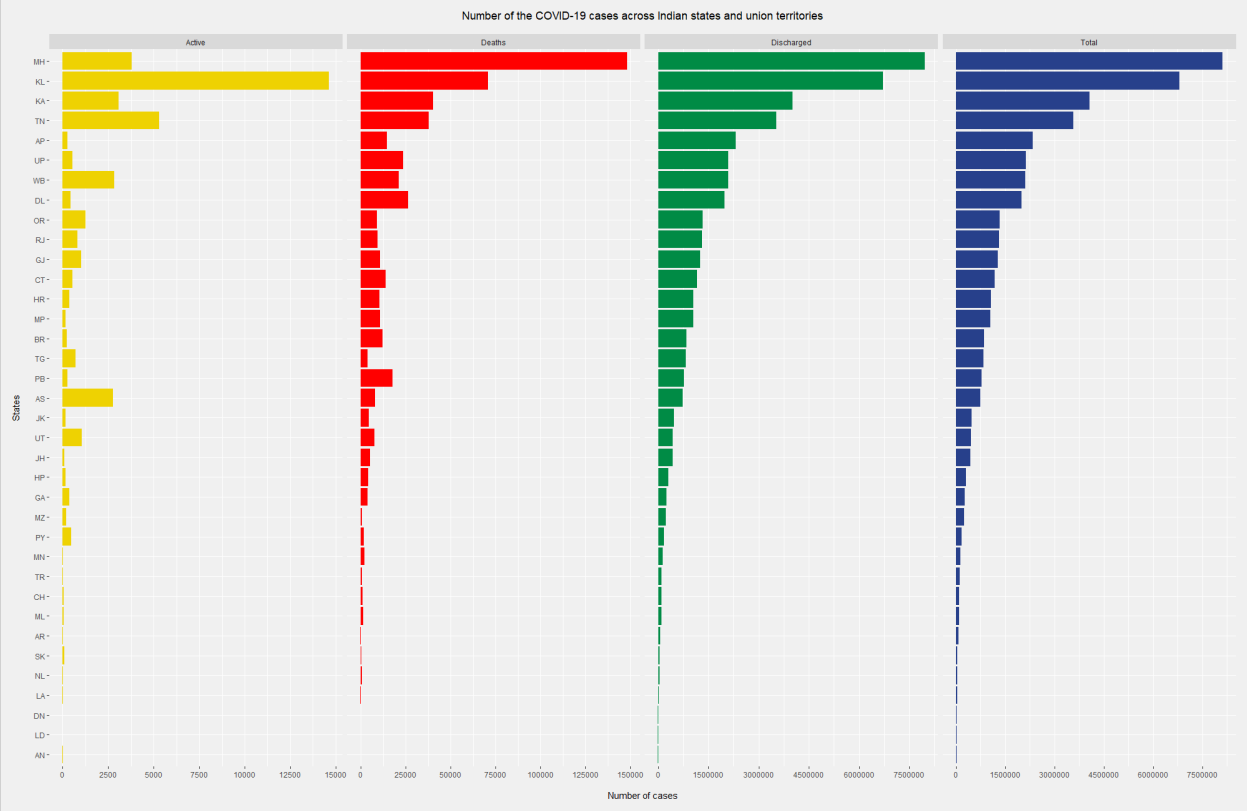
The data was sourced from the government website where they had a huge range of covid data for all the states which were being dynamically imported. The data is time scaling so additional data that was released after the visualization was created was used in the creation of the redesign.

Data preparation:

The data imported was accurately serving the needs of a normal visualization but to use facet wrap we had to melt the data frame into value and variable pairs with state names as the id variable. The melted data was plotted individually based on the category of the variable.

Graph design

The above bad graph can be improved by showing the individual bar graphs for the total number of cases, active, discharged, and deceased cases. In addition to that states can be sorted based on the total number of cases for better visualization. By doing this, the observer can easily understand what states experience the highest and lowest number of total, active, deceased, and death cases. The re-designed graph is:



The above graph is the updated design, it shows the separate graphs for the total number of cases, active cases, deceased cases, and discharged cases with the respective scale on the x-axis. And the states are sorted based on the total number of cases.

2.2.4 Challenge

The most challenging part of this graph was that the source indicated had data hosted onto its website dynamically through APIs. The API was then identified, and the JSON response of the API was pulled and stored which was then converted into excel format and further cleaned by removing unnecessary columns and errors occurred while data pulling as JSON array format.

The graphs were also plotted using facetwrap which made the scale customization arduous because the data was different in orders of magnitudes of 100's, we had to write a user defined function to execute the logic of a customized interval scale for each graph in the facetwrap group. The coloring scheme for each graph was also a challenge.

REFERENCES

[1] R. C. Team, "R: A language and environment for statistical," R Foundation for Statistical Computing, Vienna, Austria., 2013. [Online]. Available: http://www.R-project.org/. [Accessed 10 october 2022].

[2] H. Wickham, M. Averick, J. Bryan, W. Chang, L. McGowan, R. François, G. Grolemund, A. Hayes, L. Henry, J. Hester, M. Kuhn, T. Pedersen, E. Miller, S. Bache, K. Müller, J. Ooms, D. Robinson, D. Seidel, V. Spinu, K. akahashi, D. Vaughan, C. Wilke, K. Woo and H. Yutani, "Welcome to the tidyverse," *Journal of Open Source Software,* vol. 4, no. 43, p. 1686, 2019.

[3] B. Auguie, "gridExtra: Miscellaneous Functions for "Grid" Graphics," R package version 2.0.0, 2015. [Online]. Available: http://CRAN.R-project.org/package=gridExtra. [Accessed 10 10 2022].

[4] Plotly Technologies Inc., "Collaborative data science," Plotly Technologies Inc., 2015. [Online]. Available: https://plot.ly. [Accessed 10 10 2022].

[5] M. Hemant, "twitter," 22 April 2019. [Online]. Available: https://twitter.com/MohapatraHemant/status/1120212609045655552/photo/1. [Accessed 10 October 2022].

[6] J. Desjardins, "Chart: The World's Largest 10 Economies in 2030," Visual Capitalist, 11 January 2019. [Online]. Available: https://www.visualcapitalist.com/worlds-largest-10-economies-2030/. [Accessed 10 October 2022].

[7] S. Kanwal, "Number of the coronavirus (COVID-19) cases across Indian states and union territories as of September 1, 2022, by type," Statista, 1 September 2022. [Online]. Available: https://www.statista.com/statistics/1103458/india-novel-coronavirus-covid-19-cases-by-state/. [Accessed 10 October 2022].

[8] National Payment Corporation of India, "UPI Product Statistics," National Payment Corporation of India, 2022. [Online]. Available: https://www.npci.org.in/what-we-do/upi/product-statistics. [Accessed 10 October 2022].

[9] International Monetary Fund, "WORLD ECONOMIC AND FINANCIAL SURVEYS, World Economic Outlook Database," International Monetary Fund, 2022. [Online]. Available: https://www.imf.org/en/Publications/WEO/weo-database/2017/October/download-entire-database. [Accessed 10 October 2022].

[10] S. Ireland, "Standard Chartered: By 2030, These 10 Economies Will Be The World's Largest," CEOWORLD magazine, 8 January 2019. [Online]. Available: https://ceoworld.biz/2019/01/08/by-2030-these-10-economies-will-be-the-worlds-largest/. [Accessed 10 October 2022].

[11] G. o. India, "#IndiaFightsCorona COVID-19," 2022. [Online]. Available: https://www.mygov.in/covid-19/. [Accessed 10 October 2022].