Visualising the impact of amount invested by Government on roads vs the Death count caused by Road Accidents under varying unemployment and car sales.

Abstract

Road accidents are one of the major causes of unexpected deaths. Although some crashes are minor, some are gruesome and lead to horrific results. The cause of road accidents is often linked to poor quality of roads. This is because roads bear a lot of wear and tear all year round- under changing environmental conditions and traffic flowing across all the time. The result is damaged roads, often with unevenness or potholes- a big cause of road accidents. Governments across the globe, have realised the importance of good roads, and invest a huge chunk of their budgets each year on their development and maintenance.

Hence there is a need to study about what impact, if any, the huge amount invested by Governments has on the accidents that lead to deaths.
We select 3 countries-Canada, France and Australia, because they are among the top 10 countries that have the *highest number of highways in the world* [1].
We find out the quartile in which each Death Count of these countries falls compared to that of 53 countries each year. We also include variations in 'Unemployment' and 'Percentage change of the number of Car Registrations' in our study. The study is observed during a 20-year span: 1994-2013.

1. Dataset

There are in total, 4 datasets used as part of the visualization:

1. Infrastructure Investment [2]
   This dataset gives an insight of the amount invested by a country in transport construction, like Road, Railways etc. The subset of data downloaded was for amount invested on Roads, for the years 1994-2013. The shape of the datasheet is: 927 rows, 8 columns.

2. Road Accidents and Death Count [3]
   This dataset contains information on number of individuals who were involved in road accidents and died, either immediately or within a month of the accident, excluding Suicides. The values are measured per million inhabitants of the country, per year. The measure is inclusive accidents involving buses, coaches, trolleys and other road vehicles, used to transport goods or people. The downlaoded dataset has 1013 rows And 7 columns.

3. Passenger Car Registrations [4]
   This dataset refers to the number of newly registered passenger cars/vehicles, either private or commercial. Data is represented as growth over previous period. It hosts 725 rows and 7 columns.

4. Unemployment (% of total labour force) [5]
   It shows us the unemployment ratio as percentage of total labour force with respect to country and year. The shape of this dataset is 269 rows 64 columns and hosts data of the years 1960-2019.

The characteristics of Big data that have been used are:

1. Variety: Since there are different sources of data that are integrated. Out of the total 4 datasets, 3 are merged into 1 comprehensive dataset by implementing inner join. Variety in the project is seen because there are a lot of data categories that are taken into account-

   Note that all of the below are considered for different countries, over a span of 20 years(1994-2003):

   a. The amount invested by Government on roads
   b. The number of deaths caused by road accidents
   c. Percentage Change in the number of Car Registrations over the years
   d. Percentage of Unemployment over the years
   e. Quartile in which the Death Count fell during the country’s 20 years’ span
   f. Quartile in which the Death Count fell, as compared to values of 53 countries, for that particular year

2. Volume: The data sets are huge. For a major part of the project, the subset of data that caters to our project scope is taken into account- Australia, Canada and France, for the selected years: 1994 - 2013. But for some aspects and features, for instance, calculating the quartile in which the accident value of any country belonged to, in that particular year, the whole dataset has been used, which has 1013 rows. The quartile is represented in the form of color variations among the data points.

2. Data Exploration, Processing, Cleaning and/or Integration

The 4 CSV files are read one by one from our local file system using Pandas library in Python.

Since the values of amount invested by the Government is in Billions, the value by default gets converted to exponential form. The column is converted to float to prevent this. It is also divided by ‘1,000,000,000’ to convert the values to Billions.

Three data sets- Road Investment, Road Accidents, and Passenger Car Registrations are merged into a single DataFrame, on inner joins based on the country and the year of record. The column names are adjusted accordingly.

The fourth data set is taken up without merging, and is used to make a circle glyph, depicting unemployment over the years. List of countries and years is prepared through our previously merged dataframe, to retrieve unemployment attribute information for the same. Since the layout of this dataframe is different, a transpose is taken, and new column names/index is set.

For circle glyphs,

Size scheme is of 2 types:

1. Contsant size, as in Actual Data Points
2. Varying size, according to the values, for example, in Unemployment and Percentage Car Registrations. For this, Min Max Scaling[6] is done to bring values between 0 and 1. The new value is then multiplied by 25 so as to generate large circles for large values, and a constant value, 4, is added to each, to make sure that even the scaled value 0 gets plotted with the minimum visible size 4. This is done to 2 columns:

Unemployment Value and Car Registration percentage value. Since the Car Registration Percentage also contains negative percentages, the negative values are converted to positive, prior to scaling, to just bring out the magnitude of the percentage change. The negative values are taken care of by the colors.

Color scheme is as follows:

1. Within a country:
   Each value is compared to the country's 20 year values, and a quartile range is made, defining 4 quadrants, and assigning a different color to each quadrant's points-
   Black – Fourth Quartile (Highest Values)
   Red - Third Quartile
   Orange - Second Quartile
   Pink - First Quartile (Lowest Values)

2. Between O.E.C.D. countries:
   Each value is compared to all 53 countries' value points for that particular year, and a quartile range is made, defining 4 quadrants, and assigning a different color to each quadrant's points-
   Black – Fourth Quartile (Highest 25 Percentile)
   Red - Third Quartile
   Orange - Second Quartile
   Pink - First Quartile (Lowest 25 Percentile)

3. Positive/Negative:
   As in Percentage Car Registrations, before MinMax Scaling, a new Panda series is made by applying a lambda function to the percentage value columns, to assign a 'Red' color for negative values, and the 'Green' color for positive ones.

Apart from these, a green regression line is rendered, according to the best fit among the actual data points. The slope and intercept of the line is calculated by implementing the 'polyfit' method of Numpy library.

3. Visualisation

To represent a relationship between the Death Count and Amount Invested on Roads by Government over the years, an interactive line chart is rendered with many different glyphs and markers to bring out extra information and relationships.

First, all the graphs are separated using tabs in a panel, with each tab representing a country to switch between- AUStralia, CANada, FRAnce.
In each graph, top right corner shows options, for different glyphs- so that we can see the type of information that we want to discover, while hiding others.

Information that can be revealed from the graph:
In the graph, lines and circle glyphs are present, of varying sizes and colors, the background of which is as follows:

1. Actual Data Points (*Circles*)
   The size of the circles is constant throughout.
   The color of the circles represents the quartile in which the Death count of that country in that particular year lied in, when compared against the values of 53 countries for that particular year.
   The color scheme is as follows:
   Black – Fourth Quartile (Highest Deaths)
   Red - Third Quartile
   Orange - Second Quartile
   Pink - First Quartile (Lowest Deaths)
   This color scheme is set because all these colors are easily distinguishable and with increasing intensity of the color (Black>Red>Orange>Pink), the attributes increase in actual value.
2. Actual (*Line)*
   A blue line passes through the data points in order to depict the actual flow.
3. Trend Detection (*Line*)
   A regression- 'best-fit' line passes through the graph, predicting the overall trend.
4. Car Registrations- Percentage change (*Circles*)
   The size represents the magnitude.
   While a circle with color Red implies a decrease in Percentage of Car Registrations over the previous year, a Green circle implies an increase.
   This scheme is set so that the viewer can infer an increase or decrease effortlessly.
5. Unemployment (*Circles*)
   The size represents the magnitude
   The color represents the Quartile in which each Death Count falls in the country's 20 years' span- Black(Top 25%), Red, Orange, Pink(Lowest 25%).
   This is done so that the user can study the unemployment trend while in various phases/quartiles of the Death Count, and to verify if a relationship exists within the two.
6. Panel of Tabs
   Tabs are present to switch between data/graphs of the countries- Australia, Canada and France.

Below are the tools[7] the visualization is powered with:
(All the tools can be toggled on or off from the toolbar)

1.Selection tools- Lasso select, Box select, Poly select and Tap
Purpose: To select a portion of graph and focus upon it, while dimming others.

2. Pan -
Purpose: To move/scroll through the graph.

3. Hover-
Purpose: To display additional data when we hover over the data points- Year, Amount spent

4. Zoom tools- Box Zoom, Wheel Zoom
Purpose: To zoom over congested data points

5. Reset-

Purpose: Exploring the graph with the help of these tools can often turn out to be messy. We can reset the changes and bring the graph back to the original shape and form with the ‘Reset’ tool.

All this was possible using the following libraries of Python:
Math
Numpy
Pandas
Bokeh

4. Conclusion

Roads make travel & connectivity easier, transportation faster, and hence price of goods also comes down, leading to more profit and comfort. They also contribute to the economic growth of the country. It is no coincidence that countries with the highest GDP have the best quality roads[8] .

We see from the graph that clearly a correlation exists between the amount invested on roads and the number of deaths due to road accidents. Over the years, with varying unemployment rate and car sales, the death count on roads has only gone down. The downward trend can easily be noticed in each of the 3 countries. Governments have started to spend huge amount of money, unlike before, on the development and maintenance of roads.

Some points worth noting in the visualisation:

1. Owing to the Great Recession of 2008 [9], the Percentage change in Car Registrations have big Red Circles for 2008 and 2009, due to lesser sales of cars in these years.,also leading to lower deaths.
2. France has experienced major downfall in the number of road accidents. While 20 years back it had its place in the Top 25 Percentile Countries with road deaths, it managed to make its way to lowest 25 Percentile in 2013. Also, French Government has spent more amount on roads the Australia and Canada, which can easily be observed from the visualisation.

However, we should always note that correlation does not always mean causation[10]. The decrease of number of deaths might also be affected due to other factors- over the years quality of the cars have increased and people now stress upon better quality cars. Since this is an unquantifiable measure, it could not be taken into account. Secondly, over the years things like road signs, better and reliable traffic lights have become more common unlike 20 years back. Also, attaining a driving licence has become more difficult than ever, leading to a higher percentage of skilled drivers on the road. The number of road accidents could also be affected by these factors.

References:

[1] Wikipedia, “List of countries with the highest number of roads.” .
[2] “OECD (2019), Infrastructure investment (indicator). doi: 10.1787/b06ce3ad-en (Accessed on 16 November 2019).”

[3] “OECD (2019), Road accidents (indicator). doi: 10.1787/2fe1b899-en (Accessed on 17 November 2019).” .

[4] “OECD (2019), Passenger car registrations (indicator). doi: 10.1787/c58fcf22-en (Accessed on 24 November 2019).”

[5] The World Bank, “Unemployment, total (% of total labor force) (modeled ILO estimate).”

[6] R. DeFilippi, “MinMax Scaling,” Apr. 2018.

[7] “Bokeh Docs.” .

[8] International Transport Forum, “Road Safety vs. Economic Growth.”, pp. 18-23, 2015 .

[9] “Great Recession.”

[10] A. Kelleher, “Correlation Doesn’t Imply Causation,” Jun. 2016.