Good afternoon everyone

Welcome to my presentation

My name is Avinash Prabhakaran.

I am currently studying Master of Data Science in the University of British Columbia.

So the topic i have chosen today for presentation is data visualization.

(Over view) - How to show the structure of the presentation ?

I am sure that all of you would have used data visualization at some point in your career. And if not surely you would have dealt with it in you school or college days.

+

So what is data visualization ? It is a **visual representation of data** for the purpose of making a decision and communicating message and understanding your data.

I am going to concentrate on the role of data visualisation in the field of data science. Here you can see the structure of a data science model. So once we have the data in the format that we want( which is the tidying and transforming part. Then we can visualise it.

Like any other project we will have to present our findings and results. and data visualization is going to be your friend here. Because it is always easier for people to understand data using graphs and images.

Data Visualization tools

So there are a lot of good data visualisation tools out there such as PowerBI, matplotlib, D3. and so on. But for this presentation is am going to talk about ggplot2.

Introduction to ggplot 2

R is a statistical programming language and is heavily being used in the field of data science due to the amount of statistics involved with it.

??? grammer of graphics -> talk a few words about it

And there are different ways you can use ggplot. If you want to just want to plot basic graphs then they have a fucntion called qplot which is a very quick and easy method to use ggplot.

If you want are very specific about the type of graphs you want to plot and want to access the advanced features of the library then you have to use the function ggplot. So all the graphs you will be seeing later is drawn using ggplot.

Since ggplot became so successful in the world of data science that it has also been transferred to python.

Why ggplot2 ?

Here you can see some of the reasons why ggplot2 became so successful.

It is free and open source - SO you do not have to pay anything to use it. And you have access to all the codes for the library and if you feel like contributing to the library there is that option also available for you.

Professional quality graphs - Being free is not the only consideration when people choose a tool. The quality of the graphs produced is also very important

Then ggplot2 has a literate programming style - So basically you will be able to read the code and understand what you are doing with each step. And you will be surprised how transparent the code is!

When I say superposition what i mean is .. if you want to have bar plot and a frequency plot within a single graph then ggplot can do that for you.

So now that you have created your graphs. You want to customise it and make it look even better that it already is. They have a many inbuilt themes that you can easily use to customise the plots. And if you are still not satisfied with the themes that are available to you. You can access a lot more themes using the library ggthemes.

Similar to ggthemes there are a lot of libraries out there for ggplot that can that extend the capabilites of ggplot2

And finally the most important reason why i love ggplot2 is tidyverse. They have these amazing packages that can easily let you get the data you want

Tidyverse is a set of libraries that aid

Tidyverse is a group of libraries in which ggplot2 is also part of. .Before you can actually visualise your data you will have to slightly modify it. And this is where tidyverse comes in. These libraries are so easy to use that it will put a smile on your face when you will use it. It is that easy and powerful.

So before we get into the data visualisations, let us quickly go through the dataset I will be using to show you the power of gapminder.

SO this is just a few rows from the gapmidner data set . There are a total of 1704 records. In the dataset there are XX countries and for each countries there are 12 records. starting from 1952 to 2007. There are records every 5 years. And there are 3 main variables in the dataset. Life expectancy of the country, the country’s population and the country’s gdp per capita. Gdp per capita is just the total gdp of the country divided by its population.

Box plot

So here is the first plot. So what i have done here is the use a box plot to show the distribution of gdp per capita for all the countries across continents in 1987. From a box-plot you will get an idea of how the gdp is distributed for each of these continents.

Transition to jitter

Well box-plot is well and good. it gives you a lot of information like the 25th percentile, 50 percentile and 75th percentile and so on. But this can still get better.

Jitter Plot

I can actually superposition a jitter plot on the box plot to show actual data points used in plotting the box plot. This can be very useful if the audience wants to know more information about the data or if you want to know more about the data yourself.

Geographical plot

So this is a geographical plot of the world. Where the colour of each country gives us information about the country’s gdp per capita. For creating this plot i have also used another data-frame to obtain all the countrys’ latitude and longitude. And the countries which are represented by grey are the ones for which the gapminder does not have any gdp information.

Contribution of GDP

So now i want to see how much a country contributes to a continents gdp. So here we can see a stacked bar plot where area of each section gives the contribution of gdp to that continent by that country. So if you look at americas you can see that America is the major contributor to Americas followed by all countries other than the ones specified here. And they are positioned according to the proportion of their contribution.

So now lets move on to exploring another variable.

This is a violin plot. It is similar to a box plot but it shows the kernel density estimate. What that means is the width of the plot is determined by the number of data points in that region.

And similar to what we did with the box-plot, we can add a jitter plot on top of the violin plot

Say that i want to explore the variation in life expectancy across the years for all the countries. This is what i would do. So I have created a line plot for all the countries and then faceted by continents. So facet is just each country having its own sub graph.

So from this graph i can see there there is some variation in life expectancy in Africa and Asia. But since Africa has the greatest variation , let us explore the life expectancy in Africa at greater depth.

### Shift polar bar plot up

Polar bar plot. So this is a bar plot in a a polar coordinate system. And it is very simple to create a polar bar-plot from a normal bar plot. It is just addition of a single line of code.

So this a facet graph of life expectancy in africa in 1952 and in 2007. So by the looks of things there appears to be good improvement. But this is not very helpful in revealing the variation in the data.

So this is just a normal bar plot but i have played around with the data a bit to obtain the change in life expectancy in some of the African countries from 1952 to 2007. And the results were pretty shocking to me as i did not expect any country to have done worse that itself from 50 years go. Even though this does give us more insight into the African life expectancy, let us keep digging deeper for more information.

So this is line and point plot for some of the selected countries that i wanted to have a closer look at. I have also added Canada for a reference. And the difference in startling . It is massive. If we want we can dig deeper into the reasons why there is a sudden dip in the life expectancy for some of the countries given here. But due to time restrictions i will not go into that.

So no that i have explored the life expectancy and gdp per capita, I would like to see if there is a relationship between the two. So this is another facet plot. So what i have done here is a scatter plot between life expectancy and gdp per capita but the size of each point is determined by its population. So here you can see, the two giants in Asia are India and china. From a quick look i think there is a small increase in life expectancy as gdp per capita increases.

So this is scatter plot with for all the countries for all the years from 1952 to 2007 and when i fit a linear regression model into this I can see that there is an increase. So this is just a linear regression model that i have fit into the data points. So if you want to have a more complex model, then it will look something like this. And plotting these models are super easy. It is just two words. and changing the model from linear to non linear is just another argument. If you would like to have a look at the code then once i complete the presentation I would be more than happy to walk you through what i have done.

So that is it with plots. I have gone through some of the main plots of ggplot and tried to show you how flexible and easy to customise it is!

Now onto the last section of the presentation.

ggplot2 Extensions

So these are a few packages that can be used to extend the capabilities of ggpllot.

The first one is ggmap. It can be used to download a map and then we can modify it or plot points on top of it like i have done here.

The next one is ggnet2. This is a network graph and these can be very useful when you want to represent some kind of network. An example can be a graph to show the connection between a group of people on facebook or twitter. So what you see here is an undirected graph. You have to option to make these graphs directed as well as allocate weight to each of those nodes if required.

And next is Plotly. It is a 3D plotting library. It can be used to produce very visually appealing interactive 3D graphs. Since i could not embed the HTML