

**Advanced Financial Instruments for Sustainable Business and Decentralized
Markets**
Prof. Abhinava Tripathi
Department of Management Sciences
Indian Institute of Technology, Kanpur
Week 9
Lesson 27

In this lesson, we will perform basic operations related to data exploration. First, we will read all the price series data. These are carbon market that is European Union and launch features, ESG fund, S&P 500 and Bitcoin data. Next, we will visualize the data through various approaches to understand the basic characteristics of the data.

In this video, we start our session by loading the relevant libraries, then reading the data and setting the working directory. So we will start our session on portfolio analytics.

For this particular session, the most important library here is portfolio analytics library. We will load that and then another very important library that we are going to use is tidyverse. This is an advanced library for data analytics. Lastly, we will also load library patchwork.

So now we have loaded the relevant libraries, load the relevant library packages. Next, we will set the working directory. If you are familiar with R, this is a very important step. In future, whenever you are working on this file, you need not remember the location where you have saved the data and all by setting this working directory. Look at this set working directory, choose directory. Then in this particular location, I have all my data files. So I will set it like this and for record, I will copy paste this particular command and put it here. So in future, I do not have to remember the location, I simply use this address, run it with control enter and set the working directory. Whether the working directory has been set appropriately, I can see on this file command. All my relevant files for this session are visible here and I can use them.

In particular, two files I am going to use in this session. First I am going to use asset price data which is in this asset price XTS RDS file. When I click on it, a command has been run. Those who are new to R, you can just copy paste this command. So in future, if you need to read a data file, you can straight away go to this and run this command. This arrow symbol is the same as equal symbol. Assignment is called assignment operation. You can run this and you will have the data file. If you want to have a brief look at this, see the class. This is a Zoo object.

You can run a summary command on this. So if you put a summary command on this and before doing summary command, let us do the head command. So we know what are the elements here. So since I have already worked with this data, this contains the prices and

year value. So it contains the prices of Bitcoin, S&P 500, ESG fund and EUA futures.

EUA futures here is the carbon market which we have studied. ESG fund is the one of the ESG, major ESG fund and S&P 500 is a major index in US and this is Bitcoin data. Along with that I have put the year here and the date is marked as timestamp because it is a Zoo object. Date is not a variable, it is just timestamp to all the observations. This is a unique feature of Zoo object and we will make use of it.

You can also run the summary of this object to get the summary of prices. That is very easy but we will have more idea about this with visualization. Also you can see the structure of this object with this structure command. You can see what are the objects. You can see it is an XTS object with the number of elements.

See there are 753 rows and 5 columns. These columns include as we saw one year column and four data columns. You can see their properties very well in detail here. There are more such commands where you can examine your data. Starting with, let me show you some of the very good and important commands.

You can check the dimensions with Dim command. You can see the dimensions of, this is often called a data frame also. This XTS object is in a data frame form. So you can see there are 750 rows that means 753 daily observations and 5 columns. You can also run this glimpse command, a very useful command which is part of this tidyverse packages, part of dplyr.

You can run this glimpse command and see a brief overview of this data. Similarly, another very important aspect of this is, this analysis would be return data. So we will load the return data in a very similar manner. I am copy pasting this command also for our reference. Now in a similar fashion, the way we have run the price data, we will also use the return data and I will just copy paste these commands.

I need not type them again. So for example, I can check the class of the data. Again, it is an XTS view object. Then I can also see the head of this. You can see some initial observations.

These are returns for Bitcoin, S&P 500, ESG fund and EUA futures carbon market. So these are return observations. You can also run the summary command to see a brief summary of returns. See so we have the minimum first quantile median, third quantile and max for all the Bitcoin, ESG fund, S&P 500, EUA futures their returns. We can also check the structure of this data which will be pretty similar to the price data.

You can check the structure. It gives us all the details. One observation is less which was

consumed while making returns. So there are 752 observations. I can also confirm this with the demo object.

Lastly, I can also run this glimpse object to have a brief overview of the data available at hand. So with this, we conclude the first set of video where we loaded the data. We set the working directory, loaded the relevant packages and libraries that are to be employed. And lastly, we also did a brief summary of our data that we have loaded. Starting from the next video, we will move on to visualization session for a greater understanding of our data.

In this video, we will continue with our data visualization session. We will be using ggplot function which is one of the most advanced version of what we call as grammar of graphics. This is the most advanced version of plotting data in a very simple manner. So first, we will create a ggplot object, ggplot. We will make use of our asset prices which is asset price xds data.

Let us make it slightly more simple as asset. So I am creating a new name or a copy of this data which is asset. I do not want to disturb my original data which is a good practice. So first, we will give the name asset. Next, you need to specify the aesthetics that you want to plot.

So I want to have on x axis the dates which is extracted through this index command. Through index function, I will extract the dates from my asset object. It is a time series view object so it is easy to do this through index command. Then I want to plot the time series of bit. So in this particular video, we will see how to plot the time series of prices.

So first, we will plot it for the bit coin. So we have the object name as bit so I will use y equal to bit to specify my y axis. In the next layer, so this is my first layer, so if I run this layer, you will notice a sort of plot, empty plot is created. I need to specify what kind of plot I want which I will do in the subsequent steps. So in the next step, I will specify that I want a geometric object in the form of line command.

So I will run this line command and I will specify a color which is blue. I want a blue color line here so I will specify this. Let me see what comes out of this. So notice a red line, a blue line has been plotted here. We can see the price over time from 2018 to 21 of bitcoin in blue line form but this needs to be improved much more.

So for now, all we will do is we will add some axis also. So the axis can be added with this lapse function where I will use title as the central title as bit coin. So I am improving my chart a little bit. Also I want to keep my y axis empty so I will not use anything on var which is done simply by specifying a blank between quotes so there is nothing.

On x axis I would like to add the date. So now if I run it, a very nice plot will appear and I will save this. So this is a nice plot with all its dimensions and I will save it in the form of p1. So my p1 is created. Now for p2 I will not rewrite the full thing I will just copy paste the same command and the second graph that p2 that I am going to use is of ESG fund. So here instead of bitcoin I will have ESG fund.

I rather will use a different color. Let me use the green color here and I will specify on the heading also that this is an ESG fund. ESG fund and I will save it as p2. In the third step p3 I will use my snp 500. So I will specify it as snp 500.

I am using the exact name as in my data. So if you want to check you can just do a head command and you will find the exact name bit snp 500 the same I am using here. However for setting up the axis name I can choose as I wish. So this is just character to represent the axis so I am just putting as I find it more suitable. So this is and I would like to have a different color maybe red color.

I want to keep these colors separate. The last one is EUA futures which is the carbon market. So these are the four data points that I have used. So this is EUA futures. So I am making use of these four. EUA futures, bit coin, S&P 500 ESG fund these four dimensions and I am plotting them along the date.

So I will use here EUA futures which is essentially the carbon market data from ECX which we have already discussed. I may use a different color maybe black here as a color and this will be my p4. So I have all the plots starting from p1, p2, p3, p4 ready. Now if you notice we have initially at the beginning stored the patchwork library which will help us in drawing a very nice plot. So for example I will run this p1 plus p2 plus p3 plus p4 and it will draw very nice looking plot for us as you see.

Though it would require some further changes so you can see a very nice plot but requires some changes some improvements which we will do through this very interesting command called plot underscore annotation. Here we specify the central title for the plot which is time series of prices. So this is our main theme here. Then I need to specify a little bit about the dimension of this plot. So this is done through the theme argument where we specify the theme as plot the title.

So we wanted to set plot the title with certain dimensions. So these dimensions are element dot underscore text. This is a very useful argument element underscore text as you see. So in this element underscore text I will use first and foremost I will adjust the it is adjusted by half so it will come into the center so I am setting with this h just which is h horizontal adjustment by half so it will come in the middle I will increase the size to 20 so the heading central heading looks a little bit bigger and its font as bold. So now let me adjust the command a little bit so it is visible to you on my screen.

So now if I run this plot let me run this entire command and a very interesting plot will appear as you will see a proper nice plot where we can see the price movement of and I can zoom it for you so let me zoom it so you can see now the price movement the time series of prices for Bitcoin in blue ESG friend in green S&P 500 in red and your features in black and you can see sort of their movement over the three repeated from 18 19 19 20 and 2021 a very nice looking and very useful plot is appeared where you can do all the visual analysis of the data.

In the next step of visualization we'll see the movement of all the price series with S&P 500. So we'll visualize the S&P 500 movement S&P 500 is often referred to as the indicator of not only US but global economic movement and therefore how all the price series behave along with S&P 500 is very important will sort of conduct some kind of regression visualization how these series are moving with S&P 500 will again follow the same our same technique of ggplot and we'll use this asset price asset object we'll specify our aesthetics as on the x axis we'll always have this S&P 500 values because we are looking at the movement or regressing the series on S&P 500 on y axis let's start with the Bitcoin so this is our first step that we already always use notice that when I run when I run this command again an empty layer will be created empty layer of plot will be created but we need to do a little bit more so first and foremost we specify that we want the scattered form we'll first we'll plot the scattered form through geom point the geom point will create a scattered set of object which we'll use in red color so you notice the relationship between Bitcoin and S&P 500 is plotted in the form of this jittery red scattered dots but we'll also add a geom smooth now this geom smooth is in the form of a regression line we'll put it a color of blue so that we can differentiate it and the method is LM LM is simply for linear modeling so it will be a straight line which is the OLS regression line will be plotted so if I run this you would notice that a straight line will be run with a formula of simple regression here this $y \sim x$ you can see the regression line in blue moving from one end to other now let us improve this plot as we did earlier with the labs command so with labs we'll specify some of the axis detail for example the title of this plot is Bitcoin versus S&P 500 so this is our central title on the y-axis we have Bitcoin and on the x-axis we have S&P 500 so this will improve the dimensions and aesthetics of our plot and a very nice looking plot will appear for us and we can save it as object P1 now like we did earlier rest of the objects we will create in a similar manner so I'm not showing them all of them I'm just copy pasting the steps so in P2 we can have ESG fund on y-axis ESG fund on y-axis and I can specify the color maybe I'll I'll give it a name as ESG fund I'll let us use the same color doesn't make much of a difference while use ESG fund all that I need to do is change here ESG fund here as well so these are not necessarily the same variable names I'm giving them as I wish only that I need to be correct here in this particular because this is picked from the variable name in our data asset so it needs to be correct as it is as they are in the asset and you can check it here with this asset head asset command you can see the initial few names and if you are

coding along with me you can see here ESG fund this has to be exactly same as here and SNP which needs to be exactly same as this

Again in the next step now that we have done for ESG fund P3 I can choose the P3 as maybe EUA futures this is the last one so UA futures and in this EUA futures I can again just specify that this is EUA futures and here on the y-axis I can specify that I have EUA futures so I'm ready with my plots now I'll run these commands all the objects are created I hope that they are error free as we can see they are error free so now all I need to do is because of that very useful library patchwork I can just specify P1 plus I want P2 plus P3 and let's see if they are created so now they are sideways so if I run P1 plus P2 plus P3 they are sideways in this fashion but you can zoom them you can zoom them to have a look at the plot and you can see a very straight line kind of relationship between S&P 500 and all the other three securities that is Bitcoin ESG fund a very straight line on average positive sloping which sort of gives us some idea about this relationship as we can see here there is a positive relationship also if you want to you are not very comfortable with this kind of presentation you can put a slash which will give in the row form which will present our plots in the row form depending upon your comfort and your taste and preference you can have a look at like this also so this is also another way of representing the plots where we can see the relationship of S&P 500 with all the three securities in a very nice aesthetic manner. So in this video we visualize the data that was our second step of visualization in the next video we'll have a look at the density plot of all these securities.

In this step of data visualization we'll examine the density plots density plot examination is very important because often we have the assumption of normality or some specific distribution about data which we'll employ while modeling therefore it is good to have a visualization of data densities to see how correct or accurate we are about the assumption about the distribution so we'll make use of density plot to visualization of data and as we did earlier again we'll make use of ggplot command first we'll specify the set object then first we'll do for bitcoin notice I want to have the density visualization across different years so I'm using this factor year which will specify each year as a factor individual year as a categorical variable so each year will be used like this and then I'll also use the fill as factor fill also as factor year so I'm using the color and also as you will see the boundary lines of those density plot will be specified through different colors for different years with this command so I specified my aesthetics in the next step I'll specify that I want the geometric object in the form of density so this will specify that that I will keep the line with line with which is a useful sort of argument and then alpha is to ensure so many times it happened that there will be overlapping plot so alpha will ensure that there is a reasonable amount of transparency so even if they are overlapping we can have some idea about what is happening in the plot behind and lastly as we do we'll add the labs command to specify our title so in this case the title is Bitcoin and I'll keep my y axis as

empty x axis also as empty my aesthetic is mapped to year so I'll map my aesthetic to year and also the fill aesthetic is also mapped to year so I'll put as fill also year so we just see if we have drawn the plot correctly to get some idea yeah the plot is run correctly so now all we need to do is save this object and in a similar fashion create the same similarly create the object for other security so because this code has run properly we'll create p1 and as we know we'll then create for p2 so p2 let us have p2 as maybe EUA future so I'll use the next security as EUA futures I'll use the headings appropriately here as your features my next p3 p3 we can plot ESG fund maybe so I'll use ESG fund here instead of EUA futures in p3 I'll use ESG fund I'll specify here also that my title is ESG fund that is my p3 and partly I'm running with control enter pressing control and I am running all these command lastly p4 I have S&P 500 so I'll write S&P 500 here S&P 500 so this is my last so I'll so all you need to do is run all those p1 p2 p3 create these objects p1 p2 p3 now now that you have created these objects let us see how they appear when I print them so p1 plus p2 plus p3 so you can see I'll make some space you can see I have put all these p1 plus p2 plus p3 together if I plot them a very nice density plot will be created with all the four objects though some improvement is needed I need to improve their headings and also it up it'll appear like this but some improvements are needed with the laps command as we have been doing so we'll sorry plot annotation command rather because now we are doing the central heading and all so central heading is created with this title density plot density plot year wise so this is our central title but also we need to provide its dimensions so first dimension is theme very important dimensions to theme object will create theme and as you recall we use this plot that title and title element text we often use this element text to adjust our central heading by half we'll make it center with this 0.5 also will increase its size related to other excess labels will increase the size a little bit by practice I know by experience I know it is 20 will be will do the job and then we'll slightly highlight it in the bold font so this will do the job so let me run this and often when you run the plotting it will large a plot window and once we run the command a very nice looking plot is appearing here let's zoom it on so it appears like this it's a very nice plot where we have the density distribution of all the four security that is bitcoin EUA futures S&P 500 ag fund and we can see how they are distributed some of them are so we can see these are not necessarily normal so some of them are skewed and some of them are bimodal so prices are often considered to be long normally distributed though we can see here that this is not exactly the case the price distribution is slightly though I have plotted here in raw form and you can see there is not a very nice bell shape it gives you some broad idea that there is some skewness and all those things and you can have a good idea about your data by plotting the density plot like I have done here and in fact I have plotted them separately year wise manner in the next plot we'll have in next data visualization exercise we'll have some more idea so plotting them in the box plot to see whether there are some outliers and how data is distributed across years.

In this video we'll conclude our data visualization exercise with box plots. Box plot is a very important tool to see the distributed distribution and scatteredness of data and also identify the outliers so we'll conclude our data visualization exercise with box plots. So to begin with we'll again follow the same convention of using ggplot2 and we'll specify our aesthetics which includes individual series on y axis we have your features you can have also on x axis I find y axis to be more convenient in this case you can have x also have your features so fill again I am using factor year so that factor will convert year into a categorical variable then you need to tell R that what kind of geometric object you want so in this case we know that we want box plot so very easily we get this box plot exercise and then let me show you what exactly we get out of this so when I run this command we get this kind of object so we have three years on the y axis we have your features distributed for three years in different colors for example 2018 is slightly reddish color green for 2019 blue for 2020 let me improve this further before we discuss this a little bit as we usually we do we add this labs command to fill and notice this aesthetic is mapped as factor year so I need to specify this fill should be mapped as year so this is nicely printed as year and also the title as EUA futures and I'll specify the y axis we can clean it so now if I enlarge the plot let me enlarge it and a very nice plot will appear so you can see the plot here being printed.

Now this is a box plot for three different box plots are there the lower the base of this box plot is the first one tile or 25 percentile the central line of this box is the median observation so this represents the median the top is 75 percentile 75 percentile so between 25 to 75 percentile almost 50% of the observations are here the whiskers that are here these are the whiskers and the whiskers beyond the whiskers if there are any they are called outliers there is a certain formula to calculate the length of these whiskers which is nothing but starting from this 75% if you add 1.5 times what we call as interquartile range what is interquartile range IQR. IQR is nothing let me write it in the form of comments IQR is the difference between 75 percentile line minus 25 percentile line so this difference IQR when it is added when it is added on the top and subtracted from the bottom these represents the whisker length so whisker length is what 1.5 times the interquartile range and any observation beyond these whiskers is called extreme outliers that vitiate our estimation so this is how you interpret the box plot the base is 25 percentile the top is 75 the line inside at the center is the median line which represents 50 percentile the center of the data and beyond that you have whiskers which are driven by a certain formula any points beyond these whiskers are extreme outliers so as we do often we will assign this to our plot object P1 and we will get all four objects for all our data so this was first object.

Our second object would be we can take probably bitcoin so let me use bit is the name of the variable for our bitcoin data so this is our P2. Similarly we can run P3 P3 probably we can use ESG fund maybe so I will use ESG fund here ESG fund ESG fund lastly P4 we

can have a S&P 500 data so on P4 I can have S&P 500 data here we can write S&P 500 I will run all of them so so I have all the plots and let me see if I have them run error free yes I have so we have done them successfully so now first I show you the plots in the raw form before improving them aesthetically so let's see if I run this P1 plus P2 plus P3 with the patchwork library that we have installed earlier a very nice plot will appear here so you can see all the plots you can see the sort of if you look at here 2020 generally the prices are on the higher side so which in shows the gradual increase in prices that's why the distribution of the blue which is for 2020 slightly on the higher side and 2019 is on the lower side also you can see the observations of blue one are slightly more scattered which indicates the effect of covid because there are a lot of volatility and fluctuations in prices due to covid that is appearing in the scattered nature of the diagram particularly the outliers for bitcoin you can see a lot of outliers here and in the ESG fund also there are lot of outliers which indicate the covid driven fluctuations.

But let us improve our diagram further with that plot underscore annotation command so here we will specify the title as box plot year wise then we need to specify our theme object and we have seen that we have done that multiple times now so we specify the theme first we specify the plot dot title with element underscore text object and here first we put the adjustment horizontal adjustment adjust equal to 0.5 we also add size we increase the size to 20 and face to bold so we have a very nice looking diagram right coming right away for us I will run this so with this layer by layer we have a created very nice looking diagram here for all the four securities box plot you can put more focus here in the absence of time we are not doing it here but you can just spend some time looking at this graph have a look at this see what sense you can make or what you can infer about the distribution of observations I have already explained how to interpret a box plot so you can compare not only across years but also across securities for example you will find that if you look at bitcoin there is lot of scattered nature which indicates there is lot of volatility particularly in the year 2020 in fact in the year 2020 the blue one there is volatility across all the securities because of covid so all those things you can have a careful look at the diagrams and infer about the security prices.

In this video we examined the box plot and we concluded our data visualization exercise from the next set of videos onwards we will perform the portfolio analytics with the help of portfolio analytics package we will do the portfolio optimization and portfolio construction with returns data. To summarize this lesson we performed the following tasks first we loaded the relevant libraries then we set the working directory we started by reading the data from the appropriate location already set as our working directory then we checked the class of the data initial few rows and also examined the summary structure and dimensions of the data. Next we visualize the data with ggplot for each price series including EUA futures, bitcoin, S&P 500 and ESG fund data.

Next we visualize the linear relationship of all the three series that is carbon market, ESG fund and Bitcoin with S&P 500. Next we examine the distribution of each series with the help of density plots. We concluded the visualization exercise of the data with box plots. Thank you.