Week 7 Lecture 1 Case Study

Welcome to week seven. And this is case study focused on the cause of death in America, and analyses the leading cause of death in the United States of America between 1999 and 2017. This case study will try to answer the question such as, what were the cause of the deaths in this dataset? What was the total number of deaths? What is the number of deaths each year? Which can state had the highest number of this overall?

What were the top cause of deaths in the United States during this period? And also the other equations? Such as, is there any correlation among data set? And how can they

be closers to get in? So the first decision of the data analysis is data gathering. So the question is, how you're going to gather data is important together, a study data set from a reliable source is also important to use an updated and accurate data set to get on boils finding the data set in this case should be come from the open data from the US government Centre for disease control, which can be accessed to this

link. It's a co.cdc.gov, you can load and clean the data set. In order to load them, you have to use CVS from handles, which is shown here. You see CD s, from pandas in order to load the data set, and this is the exact

address of the data set that you're going to use when you're printing the data set and shape shape shows the size of the data set. And therefore, you can choose how many records exist. So what is it what is the total of the number of the recording data set,

as a part of the data gathering, we also have to clean the data basically means that we have to remove all rows with the unknown causes might have a not a number or non or an A not applicable. And we can do this using drop in a form pandas. Then you can view and print top 10 rows using dot head 10 command which is shown here. And we can look at the name of each column. So this is important because you're going to use that later on your few important columns lucky you caused me a state and that's all now we can solve it analysis or stage four of the data lifecycle data analysis lifecycle we cannot answer a question for example, how many calls exist to do so we can create an array from the field cause name using this command, and it's created a new data set called causes as you can see is an array and it has a many entry in it. Then we can create a data set of calls as a subset of the original data set. And then we can remove any extra entries such as all causes, and make every measure its lens using a subset of data. So to measure it lens, it will give us The number of the causes in this case shows that there are 11 causes is identified. We can continue analysis by recapping the number of the state in the data set for this purpose, create an array from the field state and call it a state shown here. As you can see, the output is shown here and include many assayed, but also is included extra entity like a United States. Obviously, it's not as thick. So we have to remove any extra entity, such as American state and measure the length. And we do that as before we created, we have basically every command and then we measure the length of the state by method, it gives us 51, which means that all 51 estate were included in this case study.

So 51 is included in this study, then the question is what was the total number of deaths in the United States

from 1999 to 2017, we can do that using, again, some method. And by doing that, we can see that basically, the total number of deaths during that period is over 95 million. And using the again, group by command and then song, we can see basically the yearly number of deaths for each year, which is shown in this table.

We can visualise that table. Upload the data. Rich give us a better idea about the pattern behind behind the data and help us to make the right decision in the related business.

For table visualisation, we just use plot command. And as you can see, we use le s dot plot. And in the plot argument, we put it at the title. So and this is show a plot of the table before that you've seen and as you can see the number of this decline between 2002 to 2009. And then there was a continuous growth in the number of deaths from 2010 13. Finally, chose a sharp increase in number of this since on Twitter we can now investigate which 10 states had the highest number of deaths overall. We can use grouping by state and some method and sorting them each give us basically any data set called a data set to make a use or comment here to shoot the table.

You can also use visualisation tool as the toy and to see the highest number of death estate doing that we can use again a 10 to basically focus on the top 10 estate and then plot it to produce this plot. This plot shows the California at the highest number of deaths in the United States in that period and for the coming in the second pretty close to takes us to continue table utilisation to find that top top cause of deaths by grouping data All by cars, and then solving them using some method as follows. Then we can sort it out, and we can put it as follows. And this gives us a dataset, we can, again use attend to get the top 10.

causes, we can also utilise that table to have a basically plot ball using plot for command.

And as we can see here, the heart attack has the highest cause of deaths, followed by the cancer. So, in stage five, you can evaluate and interpret the data we obtain, you can determine, for example, cause of death in America between 99 to 2017. And you can answer all the questions that were raised in the beginning, for example, what was a total number of deaths? What were the cause of the deaths in this data set? What is the number of deaths pay each year? Which state had the highest number of deaths overall? On what were the top cause of deaths in the United States and many more questions as it rises about southern Florida as a sort of noise afternoon, this United States after 10 so reflecting reflecting on your evaluation and interpretation, you may ask many more question. Will there be any correlation among data said how can we be clustered and whether you can colour correlate a sharp rise of tears in 2010 to any environmental or economical problems such as recessions, for example, and everyday you can obtain a new conclusion from them.

Your affliction in the narcissist, you would like to transfer your resolve to the actionable knowledge and therefore this may create more decisions for you to follow up. You may need new data sets such as economical growth or any other environmental disaster that might occur in that years, and to find the correlation among them. And use the classic answer or any more advanced analysis to answer the new questions. This will lead to actionable knowledge. For your design exercise, this case study will help you to analyse and visualise a data set from a research institute. You will also use advanced analytics such as ACA, which is sample Oracle clustering algorithm for clustering that dataset