* Carlo Lipizzi: It. I'm resuming the recording. Welcome everybody! Today. It's it's by day. It's a march, the fourteenth, so 3, 14,

0:16

* Carlo Lipizzi: I mean being a mathematician. I cannot escape from that. So we have she, you our ta for this.

0:29

* Carlo Lipizzi: and she will share with us some of the criteria that she is following on a grading and give you some

0:40

* Carlo Lipizzi: comments indication directions

0:51

* Carlo Lipizzi: to you. Go ahead.

0:55

* Shiyu Yuan: Thank you, Professor. May I share my screen? And

0:57

* Shiyu Yuan: i'm sorry because I use the screen, and the screen is on top of the laptop. So sometimes you may wonder why my I is brought in this area because my square is over there just

1:01

* Shiyu Yuan: functions.

1:14

* Shiyu Yuan: Okay. So can you see my

1:20

* Shiyu Yuan: Okay? Oh.

1:24

* Shiyu Yuan: I don't know why. There's some problem when I open the

1:27

* Shiyu Yuan: files.

1:31

* Shiyu Yuan: So I need to reopen it. Now.

1:33

* Shiyu Yuan: Probably. Yeah, this one, though.

1:37

* Shiyu Yuan: Okay. So can you guys see the midterm description of the word file?

1:44

* Shiyu Yuan: Hello.

1:51

* Shiyu Yuan: I just see the pie charm.

1:52

* Shiyu Yuan: Read again about it. So let's go through the

2:03

* Shiyu Yuan: problem. Other questions, one by one. So for the first one, the requirement is the question. One has string. But

2:07

* Shiyu Yuan: typically this

2:16

* Shiyu Yuan: head of scripts return us a list, because, no matter what you append.

2:18

* Shiyu Yuan: they are, is a list. This is definitely what the this other requirement to want, so we can first around the oh, i'm sorry, Professor, because you have very dense agenda. I don't want to run the

2:26

* Shiyu Yuan: this file. but we want to fix it

2:39

* Shiyu Yuan: so because they requirement to once a string we will give it history. So they they normal way I will do is just to

2:44

* Shiyu Yuan: use John

2:56

* Shiyu Yuan: to put the elements in the list together, and according to the requirements

2:57

* Shiyu Yuan: they are, they should be 3 characters like this one.

3:04

* Shiyu Yuan: So we will add the 3 letters

3:09

* Shiyu Yuan: in that list before we join the list before we put all the events in the list into a string.

3:13

* Shiyu Yuan: So this is the this is my solution for this

3:22

* Shiyu Yuan: questions. So let's run it as required. We can put the input

3:28

* Shiyu Yuan: the

3:35

* and

3:37

* Shiyu Yuan: like this is.

3:38

* Shiyu Yuan: Does anyone have a problem with this question you can ask me now.

3:42

* Shiyu Yuan: No, okay. Let's move forward. So for the second one, I

3:49

* Shiyu Yuan: to be honest that I did not follow this wrong

3:55

* Shiyu Yuan: syntax. Oh.

3:59

* Shiyu Yuan: very well. But according to the requirements

4:01

* Shiyu Yuan: I can share with you guys, my logic on my solution to figure out this problem. So calculate to. The longer is the word of a text, and printed them with a long it to the second longest word. So first

4:06

* Shiyu Yuan: we need to open the file.

4:23

* Shiyu Yuan: I read the file. If you open the file as this format and the the file, the rate, it will come up

4:25

* Shiyu Yuan: with a stream format, but we can also use red lines. I don't know if you can distinguish what is read, and reliance that rate will result with us. I I can share with you right now.

4:34

* Shiyu Yuan: So see, the word list with data is a string.

4:53

* Shiyu Yuan: This is the format. It's a string, and if we

4:57

* Shiyu Yuan: with reliance it will be

5:02

* Shiyu Yuan: list

5:06

* Shiyu Yuan: like this, but it's still with the slash. That

5:10

* Shiyu Yuan: okay, that's for the open the file.

5:19

* Shiyu Yuan: because I open the file with the stream format, and then I' it.

5:24

* Shiyu Yuan: and the

5:30

* Shiyu Yuan: what it in the pure

5:33

* Shiyu Yuan: where list.

5:38

* Shiyu Yuan: and because we want the unique

5:40

* Shiyu Yuan: one. So we take the site and convert the site into the list.

5:43

* Shiyu Yuan: So the site here is to select the unique word in the word list, and the list here is to convert to the site into the list they have.

5:52

* Shiyu Yuan: Yeah, did I have? And we use Lambda to sort the list. and we can come up with a

6:04

* Shiyu Yuan: Well.

6:18

* Shiyu Yuan: yeah, this one. because this is

6:30

* Shiyu Yuan: what happened.

6:37

* Carlo Lipizzi: Sometimes when you run a part of the code, the reading files is not working. Okay.

6:59

* Carlo Lipizzi: but I mean you gave us an idea. You can just move to the next exercise.

7:06

* Shiyu Yuan: Okay, okay, so I will.

7:13

* Shiyu Yuan: I will share with you guys with this solution on the through the announcement. Sorry for that.

7:16

* Shiyu Yuan: and for the third one.

7:22

* Shiyu Yuan: I think

7:26

* Shiyu Yuan: the Professor provide provide us a very good solution to come up with

7:28

* Shiyu Yuan: this question. So first we want to check if the input is a a digit, or if in the lower case or uppercase, and then we'll put all the V in the word list.

7:35

* Shiyu Yuan: and

7:48

* Shiyu Yuan: when you type a a character, the file will tell you it.

7:50

* Shiyu Yuan: That is a consonant, or it is about

7:57

* Shiyu Yuan: so like this one. If I put it. J. J. Is a constant.

8:05

* Shiyu Yuan: So

8:09

* Shiyu Yuan: does Annual have problem with this 3 questions.

8:11

* Shiyu Yuan: Okay. So let's move forward to the next one, the next one. I don't think I input. Oh, I input the

8:18

* Shiyu Yuan: numpy.

8:27

* Shiyu Yuan: And the the next question is to actually, this is the second section.

8:28

* Shiyu Yuan: the

8:36

* Shiyu Yuan: fourth question to calculate the tenth list of frequent words, and the average account of the worst and the longest word and average

8:37

* Shiyu Yuan: average. What length? So first. this is the way. I open this to file because we want to remove staffwards before we calculate we do the statistics.

8:49

* Shiyu Yuan: So

9:04

* Shiyu Yuan: first I will open these 2 files.

9:06

* Shiyu Yuan: One is our target file, and the second is the stop horse file, because, of course, I use read, so I need to split it

9:10

* Shiyu Yuan: into the wordless version.

9:19

* Shiyu Yuan: So this is the stop horse, stock, or list, and the air a I trend.

9:25

* Shiyu Yuan: This is the strength version.

9:31

* Shiyu Yuan: and we want to remove stuff for us. Industry and taxes. My way is to put the string into what let's first, and then remove any item

9:34

* Shiyu Yuan: in the stop Force List that up here in the target pocket for

9:48

* Shiyu Yuan: So

9:54

* Shiyu Yuan: this this is the way to split the stream format into word list.

9:56

* Shiyu Yuan: so we can see

10:03

* Shiyu Yuan: this is the string format. This is the list format.

10:05

* and then we can remove

10:08

* Shiyu Yuan: the stop course from this where list the stop was like to the

10:11

* Shiyu Yuan: Okay. So up to that, Up to this step

10:21

* Shiyu Yuan: we have a sort of clean list. We can see the difference.

10:25

* Shiyu Yuan: The number difference between these 2 were list.

10:32

* Shiyu Yuan: and then this is to count the frequency, and this is to count the length.

10:36

* Shiyu Yuan: And so, first I I do not use the collection counter. I initiate a counter dictionary to count the occurrence of a word. So first, if the word ha appear in the

10:44

* Shiyu Yuan: dictionary keys, the if they should the first time up here in the dictionary keys. The occurrence of this word

11:02

* Shiyu Yuan: equal to one, and if the word ever happened, appeared in the dictionary keys.

11:12

* Shiyu Yuan: the apparent the frequency plus one.

11:18

* Shiyu Yuan: So this is the counter

11:29

* Shiyu Yuan: we calculated from the data

11:32

* Shiyu Yuan: I have to remove. Stop, stop.

11:36

* but according to the requirement, we want the 10 min to frequent word. So we need to search the dictionary

11:39

* Shiyu Yuan: and

11:51

* Shiyu Yuan: select the top 10 worse. You don't need to worry if you have different order of the to list the frequent work.

11:53

* Shiyu Yuan: Once you follow the procedure, the procedure, or the similar procedure.

12:05

* Shiyu Yuan: they all that doesn't matter.

12:12

* Shiyu Yuan: And next, we want to count the average account of the world. So we

12:15

* Shiyu Yuan: took the dictionary value which is their current. Yeah, which is the currents of the unique words. And we technique of that

12:21

* Shiyu Yuan: values. So we can see the outreach

12:32

* Shiyu Yuan: frequency is 1.2 5 5 around this, and then we use the similar way to calculate a less.

12:39

* Shiyu Yuan: So first they initiate the

12:53

* Shiyu Yuan: we put the word

12:57

* Shiyu Yuan: in the list as the dictionary keys and the length of the word as the dictionary values. And then we saw the dictionary and use the similar pro the same procedure as the frequency we got the average length.

13:00

* Shiyu Yuan: Oh. they should be less. It doesn't matter.

13:18

* Shiyu Yuan: So the average length should be 7 point to yeah. Does anyone have problem with this question?

13:28

* Shiyu Yuan: Hello. Oh.

13:42

* Shiyu Yuan: because the good job.

13:44

* Shiyu Yuan: So in the next one is to use Pandas, it's basically a data analysis of how to use pandas. And I use the function.

13:47

* Shiyu Yuan: Goodbye here, which is a very concise and

13:59

* Shiyu Yuan: yeah.

14:03

* Shiyu Yuan: concise way, a simple way. So first we use Pandas to read

14:05

* Shiyu Yuan: the data file.

14:10

* Shiyu Yuan: and my habit is to job other. I use the columns. So first we need to

14:12

* Shiyu Yuan: print out the column name.

14:21

* Shiyu Yuan: So this is the column name. I recommend you guys to copy and paste like this instead of just a typing for yourself, because Sometimes there will be a small space between the names, and you may not

14:25

* Shiyu Yuan: notice that. And if that happens, you can now to drop, or you can not manipulate to the data frame animal. That's my experience. That's my license. So we first print out all the columns, and according to the requirement, we only want to the cars.

14:41

* which means the company name and the average knowledge, and they house power to average knowledge. So we drop all the other

14:57

* Shiyu Yuan: I used call ups which make our data Freemo

15:07

* Shiyu Yuan: concise. So

15:12

* Shiyu Yuan: no, I would do. Have a our data frame only have Well, it has 3 variables, the company, the house power and the average manage. And

15:15

* Shiyu Yuan: I don't think this is

15:29

* Shiyu Yuan: okay. So based on we want to the 3 cards with lowest average Malays.

15:33

* Shiyu Yuan: So we grew by the company and take the me of the average knowledge, and the one day 3 largest one.

15:39

* Shiyu Yuan: See, this is the 3 largest 3 car with

15:53

* Shiyu Yuan: largest average knowledge. This is the company name, and this

15:58

* Shiyu Yuan: this is the me of the average knowledge for each company.

16:03

* Shiyu Yuan: and the same

16:08

* Shiyu Yuan: to that smallest.

16:11

* Shiyu Yuan: These 3 are the smallest.

16:14

* Shiyu Yuan: and the the last question is 3 costs with the highest ratio has power average. So first, I need to calculate the ratio of house power to average knowledge. This is the

16:16

* Shiyu Yuan: one to calculate

16:29

* Shiyu Yuan: the ratio of them. But I want to put the put that results to the original

16:31

* Shiyu Yuan: data frame. So I need to convert. Actually, this is a right. I need to convert them into list. And

16:39

* Shiyu Yuan: at this result to the original cars data

16:47

* Shiyu Yuan: they are free. So we can see now our data frame has fall

16:55

* Shiyu Yuan: variables. This is the power to average knowledge, and the value is the ratio of each company

17:01

* Shiyu Yuan: and for their house power to average marriage.

17:10

* Shiyu Yuan: It's out, and we follow the same function to get the 3 cars with highest ratio.

17:13

* Shiyu Yuan: Okay, so

17:25

* Shiyu Yuan: that's all

17:27

* Shiyu Yuan: for the for the mid term

17:30

* Carlo Lipizzi: great to you. So stay on life for a second that just want to show the class and you a different way to do the same last exercise. So just a although curiosity. The reason why i'm going to do it is just to explain the fact that

17:34

* Carlo Lipizzi: pretty much everybody is different. One of the things that we consider when we grade the the assignments is is to notice the differences. So the assignments are individual

17:53

* Carlo Lipizzi: meaning. Each one is doing a in a different way. If you stop sharing for a second that I will share mine just 1 min.

18:08

* Carlo Lipizzi: So this, for example is what I did the for the last assignment. So

18:20

* Carlo Lipizzi: i'm in that

18:29

* Carlo Lipizzi: I did that probably in a less a rigorous way. But it's very concise. So. So I read the file into a pandas data structure. I pre I I I didn't delete anything I I I printed the the first rows, then I I print the ratio, and then the the cards with the highest ratio.

18:31

* Carlo Lipizzi: So once you have the the head that's of the of the Pandas data structure.

18:53

* Carlo Lipizzi: You may not really need to to delete those, but just use the rules that that that you need. So I mean in this case it's pretty straightforward, and and that that's what you get. So

19:04

* Carlo Lipizzi: I mean that the first rows the last rows, the Cs with the highest ratio, and you can do more intense over what is the highest value or or the lowest value.

19:17

* Carlo Lipizzi: So I mean that keep in mind that that Pandas so has a noon pie in it, meaning you can do all the operations you want with the directly with Pandas. One thing that I want to stress with you is this operation here.

19:32

* Carlo Lipizzi: So in this case I created the a. A new column in my data frame without going through creating a list. So it basically had a the column. Now the new column ratio

19:49

* Carlo Lipizzi: by calculating the ratio between horsepower and average my larger. Then I sorted it, and then I printed the the the head, the meaning, the the ones with the highest value.

20:06

* Carlo Lipizzi: So I mean, again, there is no right and wrong apart from the fact that they generate the right results or not. But this is again just to

20:16

* Carlo Lipizzi: be sure that you got the fact that that everyone is individual, everyone is different, and and if you do in a different way. Doesn't mean that that way is wrong, or it's right, and the other by by it's just a different way.

20:28

* Shiyu Yuan: Yes.

20:44

* Carlo Lipizzi: okay to you. We really thank you.

20:45

* Carlo Lipizzi: Other people in the class. Any question, an issue, any studio city?

20:51

* Thomas Poklikuha: I have one question about the homemarks.

20:57

* Thomas Poklikuha: So for one of my homeworks. I accidentally didn't submit the pi file.

21:00

* Carlo Lipizzi: and I got points taken off, but I have the screenshots. If I resubmit that pie, I sent you a message on that in canvas. Yeah, I I mean that unfortunately, canvas is not. That is Mark, meaning. Each time you resubmit that you will cancel whatever was before. Yeah. So just send us an email with the dot pie.

21:07

* Thomas Poklikuha: Okay, sounds good. Thank you.

21:32

* Carlo Lipizzi: I apologize. But unfortunately this canvas.

21:35

* Thomas Poklikuha: It's my fault first. So it's all right.

21:39

* Carlo Lipizzi: all right. Okay? Other questions for she, you

21:42

* Carlo Lipizzi: Okay. So thanks again to you, feel free to stay if you like. Are all things that you know quite well so, and feel free.

21:48

* Shiyu Yuan: No, I I love your class. I pretty enjoy that.

22:00

* Carlo Lipizzi: Thank you. Thank you.

22:03

* Carlo Lipizzi: Erez agmoni this coming Friday I will give a a workshop on a natural language processing. I will post on a linkedin the recording, if it will be available, and the slides 101.

22:12

* Carlo Lipizzi: It took me 3 full days to Redo. It completely is probably the sixth of the seventh time i'm. Giving workshops on a natural language, processing inclusive for the same conference that I mean in the past it was in a different location was in to you. But what what was the same now I had to redo everything, because so much is changing in natural language processing

22:26

* Carlo Lipizzi: that I I couldn't reuse what I had in the in the math.

22:54

* Shiyu Yuan: and Gbt for is released today. I know, I know, I know.

22:59

* Carlo Lipizzi: I mean that i'm kind of skeptical I mean I I I see the good and the bad, but that that that's a a story for another moment.

23:05

* Carlo Lipizzi: Thank you. Bye, bye.

23:16

* Carlo Lipizzi: All right. Okay. So let's move on and let me share the screen again and let me go into.

23:21

* Carlo Lipizzi: So that was the midterm, as we we all well know. So for today we will talk about a little bit of data mining a little bit of a methodology. What is data exploration? And that's the methodology. And then we'll be talking about the visualization.

23:33

* Carlo Lipizzi: So I will go relatively fast on that, because it is already almost 70'clock, and I do not want to skip the the in class assignment.

23:51

* Carlo Lipizzi: So let me go on the slides, and let me introduce the concept of a extracting knowledge from data.

24:04

* Carlo Lipizzi: It's pretty much in line with what we already

24:16

* Carlo Lipizzi: said a few times about storytelling.

24:19

* Carlo Lipizzi: So you have data. You have a problem, and you want to use the the data to work on the problem, meaning. You want to get a facts that then you can explain. But the facts you will extract them from the data.

24:25

* Carlo Lipizzi: So sometimes things are easy, sometimes are more complicated, sometimes some some basic statistics. I can do the job for you some other times. You really need to find the more complex elements that can

24:44

* Carlo Lipizzi: create a narrative for this sort you want to tell

25:03

* Carlo Lipizzi: if you consider for a second that when we talk about the correlation, so the correlation, what we use it generate the linear correlation, meaning one value is growing, the other is growing, or one is growing. The one is decreasing. That's a correlation.

25:06

* Carlo Lipizzi: and how much at what degree they degree they they increase or decrease, is how close the the variables are to be identical in that case will be one, and the correlation 150

25:22

* Carlo Lipizzi: erez agmoni, but not always. You have a a linear coronation. So there are many other ways. So I just mentioned one example. So if you go to the 150

25:39

* Carlo Lipizzi: that sometimes we use models that that are more complex than that we use a a decision. Trees we mentioned that last time in the the class when we talked about

25:49

* Carlo Lipizzi: machine learning. So those are the methods. Those are a matrix that we extract from the data, and we use as a data points to tell the story.

26:03

* Carlo Lipizzi: So let's explore a little bit more very fast. I mean it's a lot of slides, but i'm not going to spend much time on each one.

26:17

* Carlo Lipizzi: So what is

26:26

* Carlo Lipizzi: data, mining data, mining knowledge? Discovery is all about data, but it's not the only thing that is a about data. So in this case is that somehow a process used by companies to to gardener role data into useful information. That is very basic

26:27

* Carlo Lipizzi: definition, but it makes sense. So you I mean it's it's it's basic, and it's very general. But that's what we do. So why we do data mining. Now we do now because we have data, and that's the first point. The the second point. We have a a computational power, and we have tools like a python that can really help us going through the data in a

26:46

* Carlo Lipizzi: in a timely matter and get the results. So the combination of those things again, and the availability of data.

27:15

* Carlo Lipizzi: the the computational power, and the power of all the software tools like Python is making a data mining possible. But there is also a a a market reason.

27:23

* Carlo Lipizzi: So the market environment is becoming more and more competitive, and you need to find the niches where you can leverage your offer and get more satisfied customers. So what customers, or what whatever, is the goal you have?

27:37

* Carlo Lipizzi: So now there is more need, the competition. It's more.

27:54

* Carlo Lipizzi: I'm not saying that it's hard right now, but but it's definitely more specific, more targeted than it was in the past. And then the technical reasons that I was explaining before.

28:04

* Carlo Lipizzi: So what we do, generally speaking, in the data mining is to discover patterns so common behavior. So in the data that we have.

28:18

* Carlo Lipizzi: That's pretty much what gpt for that she you was mentioning, or any of those large language models are doing.

28:30

* Carlo Lipizzi: discovering patterns and a matching patterns.

28:41

* Carlo Lipizzi: They, the mining. It's kind of a multi disciplines discipline with the several components from obviously computer science information science, because at the very end that we need to write programs, but we also need. I'll do it. So that for machine learning bid up most of the time are in data basis. You need to know how to deal with the other basis

28:46

* Carlo Lipizzi: Statistics, because even if sometimes they may not be so so sophisticated, but will be useful to clean the data, to understand the data and to start initially working with the data. And then visualization. Will we talk about that a little bit later on today.

29:14

* Carlo Lipizzi: And there might be a other discipline based on the on what you are exploring, what you want to discover. If you are in finance, you may want to be an expert in finance. If you are in a health care, you want to be an expert in that care, and so on.

29:30

* Carlo Lipizzi: So sometimes, when you are talking about sociology. You know, when we are all social media, you are talking about people like you may have a sociologist, anthropologist, psychologist on board.

29:47

* Carlo Lipizzi: Obviously, when you work with few data. Life is easy. When you work with

30:03

* 7 on terabytes of data things will become more complicated.

30:11

* Carlo Lipizzi: I was working few years ago on a a large database, all the data collected from an insurance company using the black boxes in the cards.

30:16

* Carlo Lipizzi: and that those black boxes generated input for each car every few seconds. So was a large insurance company, meaning a lot of cards, and you can do the map and imagine how big the data set was.

30:30

* Carlo Lipizzi: So my model was working fine in in theory, but when I had to scale it up to the size of the data set, then issues happened.

30:50

* Carlo Lipizzi: I ended up doing a little bit more of a a parallelization

31:03

* Carlo Lipizzi: at the very end that I ended up working on cloud computing, so moving, creating a virtual machine, moving the virtual machine in the cloud, running the model and then getting the data.

31:12

* Carlo Lipizzi: But again, scalability.

31:26

* the size of the data is important.

31:30

* Carlo Lipizzi: So one of the most common questions is, what is the difference between data, mining and and.

31:35

* Carlo Lipizzi: generally speaking, statistical analysis? It's it's top down. So it's deduction, data, mining is induction, meaning that you have a few examples, and you extrapolate what what you have.

31:44

* Carlo Lipizzi: You have a a bug over points, and you say, okay, it's full of coins, so I extracted the

32:01

* Carlo Lipizzi: 20 of them, and they are 5 cents, all 5 cents tribulation. They are all 5 cents can be good, can be not good. But that's an example, that pretty basic example of the induction. So you are

32:11

* Carlo Lipizzi: assuming that because you had the a certain number of a certain type. You are established the type to the entire population you have

32:29

* Carlo Lipizzi: in statistical analysis you count how many coins you have. And then I say, okay, 30% of my coins are of this time. Sometimes you cannot really count because you have a data set that is huge. I I was working few times with social media. You have several 1 million

32:39

* Carlo Lipizzi: data points, and and I mean it is difficult to do in

33:00

* Carlo Lipizzi: so. Again, the main difference study signal analysis is is deductive data. Mining is a. In that.

33:08

* Carlo Lipizzi: We mentioned that there is a huge growth of of data. So the data we have in the last few years, so it's pretty much comparable with the all the data has ever been collected in the previous history of humankind.

33:16

* Carlo Lipizzi: So and it's growing so because we are generating content. We are generating content through social media to communication. We have a sense of Internet. Of things. It's, Matt watches another whereabouts. So all of those are continuously generating data.

33:33

* Carlo Lipizzi: Everything we have pretty much is digital now, with very few exceptions, and they are going to be 0 sooner

33:53

* Carlo Lipizzi: if it is not digitally. If it is not the analog, that means that it's the that means that that we can use it somehow.

34:04

* Carlo Lipizzi: Now, because of most of the data are in the Internet Protocol and IP for a convenience. That means that somehow they are collectable with the all the degrees of privacy. Obviously. But they are collectable for those with the right right to do so.

34:14

* Carlo Lipizzi: So we are leaving a a a process that someone is calling the thatification a meaning. Everything is becoming a based on data. So the latest chat Gpt is probably the on this trend

34:36

* Carlo Lipizzi: Chat Gpt is based on data. It's based on that whatever, or any AI could collect the from open source, and someone is saying even something that is not fully open source.

34:53

* Carlo Lipizzi: And you don't want to argue on that. So

35:07

* Carlo Lipizzi: it's a process. You go from a a row data to element, so that we give you the opportunity to take a decisions.

35:11

* Carlo Lipizzi: There are 3 main types. All the the

35:20

* Carlo Lipizzi: I think. So the data science is is doing the analytics. So the data science is doing so. You have a the descriptive analytics meaning you are applying the method to to get what's going on from the data.

35:26

* Carlo Lipizzi: and that's basically

35:46

* Carlo Lipizzi: metrics that you extract some of the metrics. Again I mentioned the correlation analysis may be very deterministic, very statistic. I'll rather not so much. And I mentioned that the decision trees and the metrics never associated to that.

35:49

* Carlo Lipizzi: Once you have a a descriptive analytics, you may want to

36:04

* Carlo Lipizzi: go predictions. So saying, okay, considering what is happening up to this point. So this is what is going to happen.

36:10

* Carlo Lipizzi: But you don't do anything that the next step is the prescriptive meaning. Okay, considering that this is what happened. And this is what is going to happen. I'm going to do that.

36:18

* Carlo Lipizzi: Think about the autonomous vehicles. So you have the autonomous vehicle assessing the situation. So what is on the road, and is the descriptive.

36:30

* Carlo Lipizzi: then the predictive? There are a pedestrians that are crossing the street. I will see them in front of me in a that much seconds. So that's predictive. So they are not actually crossing in that particular MoD. They are not in front of the car in that particular moment, and then prescriptive. I will stop.

36:40

* Carlo Lipizzi: So those are the 3 faces. Not necessarily all the analytics. So we'll have all the 3 faces, but those are the T. V. Gala, 3 faces for analytics and the 3 TV gala usages for a data science.

37:01

* Carlo Lipizzi: Considering how much, again, that data we have this notification transform the and it's keep transforming pretty much every industry.

37:19

* Carlo Lipizzi: So some industries are more reluctant. So we still have a quite a lot of projects in the data engineering with the defense industry, because some industries are more structured and they require more time to do the transformation.

37:32

* Carlo Lipizzi: Some other industry are already transformed, and they already 100% a digital, but not all of them

37:51

* Carlo Lipizzi: examples of user pretty much all the the cases where you have a a a big use of all the data you will have the possibility to in use

38:00

* Carlo Lipizzi: data science in a in a middle sense. We already mentioned that the type of data that you can have. So most of the time you have data in a structure form that can be tables. So that can be pandas. It can be data basis, so they

38:16

* Carlo Lipizzi: pretty much resemble the table with rows and columns. We are at the rows are the data points, so that you have the observations that you have for your data set, and the columns are the attributes of of those observations. So if you have the weather in a given period of time you may have the temperature, the pressure, the humidity, those are the attributes, and then you, or variables

38:34

* Carlo Lipizzi: or columns, and then you have the the rules that are the samples. So the instances, or whatever you want to call them.

39:03

* Carlo Lipizzi: That's another example. You have a the Irs checking. If someone is cheating on taxes.

39:12

* Carlo Lipizzi: and he's collecting data like. If the subject the taxpayer asked for a refund, or what is the marital status? And what is the taxable income? And then the result was cheating on taxes was not

39:22

* Carlo Lipizzi: so based on that. You you can do a prediction. Obviously no one would do a prediction or something with so many vulnerabilities, and so much at stake, with only 10 observations.

39:37

* Carlo Lipizzi: But that's pretty much a the sense. So now, if you want to create a model to predict the who is achieving who is not. You need to analyze the the variable you have, and then create the model based on the variables

39:54

* may not be. Linear most of the time is now, but that's what you have.

40:10

* Carlo Lipizzi: Obviously, if you had the same poll that you have.

40:16

* Carlo Lipizzi: or better, if you are using the sample that you have on your screen you.

40:21

* Carlo Lipizzi: I mean, if you look, the number of No is the vast majority. You are only 3. Yes, meaning the 70 is No. If you create a model saying that always No, you will be always right, or you will be 70% of the times right?

40:25

* Carlo Lipizzi: Oh, that means that that the all that it wouldn't

40:46

* Carlo Lipizzi: do a great job. But this is

40:50

* Carlo Lipizzi: something that I would use to tell you. What is the the accuracy of a model? So.

40:53

* Carlo Lipizzi: and most of the models, they drew the line at 50%, meaning that if is less than 50% doesn't really bother.

41:01

* Carlo Lipizzi: I mean, you flip the coin pretty much, if is a above 50% may be useful.

41:12

* Carlo Lipizzi: but it really depends on what is the case.

41:22

* Carlo Lipizzi: So if you have a social science predicting what human beings can do is really difficult, and getting a high accuracy, it's really difficult. If you have a an industrial production. We are all the processes are very

41:25

* Carlo Lipizzi: a well-defined and consistent in time that an accuracy over 85 95% it's reasonable

41:45

* Carlo Lipizzi: when you have something that is not on the side, and you have an accuracy that is, 99, 95 there is something wrong most of the time you have a high correlation linear correlation between the the variable that you want to predict and the variable, so that you are using to create a model.

41:55

* Carlo Lipizzi: So think about. You want to predict the weather if it's raining or not. And one of the variables in the is the number of inches of rain.

42:15

* Carlo Lipizzi: Obviously, if that number is different from 0, there is rain

42:26

* Carlo Lipizzi: erez agmoni. So those 2 variables, rain or rain, and the accumulation of water, they are highly correlated. So if you include that that variable in the model, 150,

42:32

* Carlo Lipizzi: the model will tell you things that you already knew. meaning that the model is part of less.

42:47

* Carlo Lipizzi: So when we do this type of analysis. So you want to remove all the variables that are so obvious that wouldn't add anything to your knowledge.

42:55

* Carlo Lipizzi: So let me skip that we already know that they are tribute. So we know from 5. Don't know the categorical or numerical, pretty much the to make categories

43:10

* Carlo Lipizzi: some examples of all the documents. So you have a on data. You have different documents and different words. This is a a document term occurrence table one.

43:21

* Carlo Lipizzi: So you have a document one with the with the term a team happening at 3 times a play 5 times, and so on. So this is something that is used as sometimes as a first stage to the terminology, to see how much 2 documents may be similar if they have the same words.

43:34

* Carlo Lipizzi: What's use the the same number of times? That's a red flag for a possible. Then you would do the second set, but that will be analyzing the position of all those words.

43:54

* Carlo Lipizzi: So if you have the words, so that record in December now the same number of times in the same positions. Then there could be a platform, and obviously it's never a 100. You can have a degree on the first stage and a degree on the second stage.

44:06

* Carlo Lipizzi: The other example is transaction data. So when you buy something in a store, either online or not, you have item in your

44:25

* Carlo Lipizzi: basket, and you have a a transaction, Id that are meaning what you are buying.

44:38

* Carlo Lipizzi: So you have a transaction. Id and list of items. So that's an example of transaction data.

44:45

* Carlo Lipizzi: You can have graphs. So some graphs are pretty obvious. Social media is generating graphs. The electric greed is a graph.

44:52

* Carlo Lipizzi: The the telecommunication network is a graph. But you can create graphs from text with the proper algorithm

45:05

* Carlo Lipizzi: all right

45:16

* Carlo Lipizzi: with the problem metaphor, and then the algorithm, so that's another example. Like each one of those will have a intrinsic characteristics that can give you the possibility to that, to do some types of analysis, and not that

45:18

* Carlo Lipizzi: when you do data mining, you actually do a process. So the most commonly use the methodology is what is called Chris the end. Chris DM. Stands for a cross industry, standard process for data mine.

45:34

* Carlo Lipizzi: It is composed by 6 faces. You start with the defining: what is your business? So business understanding? Why, i'm doing what i'm going to do what is the goal that they want to achieve. So that's the first step.

45:52

* Carlo Lipizzi: If you don't know what you are going to do, chances are you will fail, or you will never be satisfied. So you define the problem. They say on the it's data mining. You need to have data.

46:09

* Carlo Lipizzi: and you need to have the data that will allow you to solve the problem that you define in the in the business understanding phase.

46:23

* Carlo Lipizzi: So this face is a sort of an assessment of the data you have. So you analyze the data. You see, if the data can provide answers.

46:32

* Carlo Lipizzi: may not, because you may not have enough data, or the day that may be consistent, or there could be so many missing values that you are not going to get much a out of it.

46:44

* Carlo Lipizzi: Things like that. So

46:57

* Carlo Lipizzi: the outcome of the data understanding that can be okay. I have what I need to do. The modeling in a broad sense, or I don't have enough data, and I need the either to

46:59

* Carlo Lipizzi: tune up somehow, or the business question, or just to table it and say I cannot do it Once let's say you are okay, and that's why you have the double

47:14

* Carlo Lipizzi: are also here. Once you are okay, you go in the data preparation. So you want to be sure that the data has been cleaned enough as being the most enough as being normalized enough. So all the or you do the correlation analysis

47:27

* Carlo Lipizzi: to the move a variable, so that are, if relevant. If you have a one variable, that is the double or another one. There is no need to keep them both. So you want to remove one of the 2. That's an extreme example, but is in in that line.

47:45

* Carlo Lipizzi: So that's the data separation. Once you have the the right question.

48:04

* Carlo Lipizzi: the right data to address the question and the data in great shape. Then you start the modeling modeling means you apply. I'll go it to the data.

48:10

* Carlo Lipizzi: So the combination of all the algorithm or algorithms and data will give you the model.

48:23

* Carlo Lipizzi: Once you have the model, meaning that the algorithm for your specific case, you

48:30

* Carlo Lipizzi: then you can start the evaluation of the model. How good the the model is I mentioned before the 85, 95, or the 50% that those are the accuracy of on the model.

48:36

* Carlo Lipizzi: Once you are okay, with the valuation. But you can also be not okay. Say, okay, I was doing a bunch of things. I tried several different algorithms, but I didn't get much.

48:50

* Carlo Lipizzi: So at that point, the the evaluation, the accuracy is so low that they need to go back to the business question as saying with what you have, I can never provide you with an answer you would be happy with.

49:03

* Carlo Lipizzi: If this is not the case, meaning that if the is good enough, then you start with the point.

49:18

* Carlo Lipizzi: So that's the

49:25

* Carlo Lipizzi: I would skip the the following slides, and I would go, but you will have them, and I would go with a specific case for what we are doing.

49:29

* Carlo Lipizzi: So we are not doing modeling, because this courses on data. Exploration is is not in data mining, but they want it to give you a little bit of context in terms of

49:43

* Carlo Lipizzi: broadening the goal and going more on the mining.

49:55

* Besides the explanation.

50:04

* Carlo Lipizzi: So.

50:06

* Carlo Lipizzi: starting from the Chris DM: I created the or I adopted the the methodology to something that is more for our cases. So data exploration, instead of having a

50:07

* Carlo Lipizzi: 3, is per

50:26

* Carlo Lipizzi: DM. Is crisp de we are instead of data. Mining is data exploration.

50:27

* Carlo Lipizzi: So you and you will use this. this.

50:35

* Carlo Lipizzi: this sort of table of content as a guidance for a a future data exploration assignments that you will have, and you can use in other circumstances of that data exploration that you may have in the future.

50:41

* Carlo Lipizzi: So you start defining what are the overall project goals? What are the questions you want to answer?

50:59

* Carlo Lipizzi: And then you start. Okay, let me introduce a methodology. The methodology is called the crease. The we are. There are 4 faces, business understanding. They'd understanding and data preparation. There are the same that we mentioned before, and they the representation. So that was a

51:08

* Carlo Lipizzi: not in the the Chris DM. That was modeling an evaluation. We don't have modeling, but we have a representation. Representation can be. Visualization can be tabled so can be not added

51:27

* Carlo Lipizzi: to consider all the sort of revolution that we are experiencing with the Chat Gpt. One of the representations is basically

51:42

* Carlo Lipizzi: conversational.

51:57

* Carlo Lipizzi: So the data are a represented in terms of phrases generated by by the board.

51:59

* Carlo Lipizzi: So that's another data presentation that is still a quite a lot of room to for improvements.

52:08

* Carlo Lipizzi: But that's a possibility. So think for a second that you are an alternative on the field. So you don't have time to go that much into a charts tables. So if you have a your system telling you in plain English what you do? What is the result of the current variable, so that you are considering you would save time.

52:18

* Carlo Lipizzi: So that's why can be so relevant that this new approach based on a conversational

52:43

* Carlo Lipizzi: data to.

52:51

* Carlo Lipizzi: So

52:54

* Carlo Lipizzi: again I will go fast into those faces. So we already mentioned that when you we do your own, instead of having a general definition that you will say, okay in a business understanding, I want to determine how to reduce the charm rate in my client base.

52:55

* Carlo Lipizzi: I want to predict the weather. I want to do something. So that's what you do in business understanding. and you will define criteria to define a success or failure.

53:19

* Carlo Lipizzi: Then they'd understanding. You will collect the data, and you will critically evaluate what's going on. So if the data, it's good enough for what you have.

53:33

* Carlo Lipizzi: So you will perform an an exploratory data analysis on the data that will be for the the for mana validity of the data. So how many missing values. How many outliers, how much correlation things like that!

53:47

* Carlo Lipizzi: Then they the preparation. At this point you will do the cleaning. You would do all the preparation that you would do in a

54:06

* in your data.

54:16

* Carlo Lipizzi: and then you will do the representation. So again, most of the time would be the tables, or a visuals. Will we talk about visuals in a moment?

54:18

* Carlo Lipizzi: But that's basically in a broader sense, data representation. Then conclusions. You want to have your conclusions here, saying, based on what I have. That's the results. And then you may have a attachments.

54:29

* Carlo Lipizzi: So you don't want to create a report with the me on tables, a 1 million visualizations. You'd want to create something that is the right combination between visuals. So tables and narrative. So the narrative

54:45

* Carlo Lipizzi: plays the to explain tables, visualizations, and things like that, so you can have the narrative first, and the visualization then, or vice versa, depending on your assign. But those will go side by side.

55:05

* Carlo Lipizzi: You don't want in your rep on to have code, because that's not the point. So we are a more like a consultant, so like business consultants, but based on on data based on facts that what we do

55:21

* Carlo Lipizzi: so attachments again, they need to be read Ebola, and then to be is useful, but not essential. So what is essential will be in the the main part of the document what is not essential, but may be relevant. That will be in the attachments.

55:36

* Carlo Lipizzi: So that's basically it for this portion.

55:57

* Carlo Lipizzi: If you don't have questions.

56:05

* Carlo Lipizzi: I will jump into the next portion that will be

56:09

* Carlo Lipizzi: on a visualization. So we mentioned the visualization is an important part of presenting the data. So let's talk a little bit more about about what visualization can do for us.

56:18

* Carlo Lipizzi: Okay, I will share the screen again, and I will go, and the the last part of the

56:36

* on the lecture

56:45

* Carlo Lipizzi: before in class assignment. They will be on a visualization, and how visualization can be done with private.

56:46

* Carlo Lipizzi: This this class is pretty much the first one after is the first one after the met or my and it's the first is the first one

56:57

* Carlo Lipizzi: that is less basic than it was in the first house.

57:07

* Carlo Lipizzi: So those of you, with the experience encoding experience encoding in Python, that from this point on that they maybe

57:13

* Carlo Lipizzi: more engaged in the assignments, because it would be more on the application than just on that. I mean that learn how to use the tool.

57:24

* Carlo Lipizzi: So we will like keep talking about how to use the tool, obviously, but with the a specific application. The application for the day is visualization. The application for next class will be the text processing.

57:35

* Carlo Lipizzi: So the history of visualization is somehow related to the history of all of the computers we had. So in the seventies, in the eighties there was not much computer graphics that we can leverage on.

57:50

* Carlo Lipizzi: I. I started working in the eighties, and there there was no windows, so the the first one was a I mean a a larger distribution was the Mac.

58:06

* Carlo Lipizzi: and before the Mac it was a it was a with the subsystem that is by, and the steep jobs with the Mac.

58:21

* Carlo Lipizzi: If you don't have a graphical user interface. There is not much that that you can do. So. In the 90. In mid eighties we start in heading a graphic and user interfaces, but late eighties, not

58:34

* Carlo Lipizzi: the the computer. We are not particularly powerful meaning. You cannot do much of the rendering that you may need. So rendering is essential to have a good representation of images on your screen. If you don't have that, then graphics and visualizations will not be good.

58:50

* Carlo Lipizzi: so

59:12

* Carlo Lipizzi: it's pretty much a 2 sides approach on one side that you develop a more powerful hardware. On the other side you develop a better techniques, but you cannot develop the good techniques without the hardware, and that can support you.

59:13

* Carlo Lipizzi: So if you look at what is happening now, I i'm using a Mac that is Apple City on so in the past that we had the the gpu, the the the CPU, the central processing unit, the doing all the calculation.

59:31

* Carlo Lipizzi: And then at the certain point we move the we introduce the Gpu, the graphical processing unit to handle up the the video.

59:47

* Carlo Lipizzi: Now, what most of the companies, including apple with this apple ceiling are are blading the line dividing the to. So when you do

59:57

* Carlo Lipizzi: calculations that are more complex automatically, the operating system is using one of the 2. So the gpu that is intrinsically parallel, because it's a it's addressing the pixels on the screen, and the so intrinsically parallel. Then you use that

1:00:10

* Carlo Lipizzi: i'm not saying that these are all good, though, is good because it's more powerful. It's faster, and that's for sure on the other end. I i'm having a really bad time using some of the traditional machine learning libraries like tens of flow that is not running on my new Mac.

1:00:29

* Carlo Lipizzi: But I mean that's a a mine of things

1:00:49

* Carlo Lipizzi: again, that

1:00:53

* Carlo Lipizzi: the history of a visualization is highly correlated with the with the history of the hardware. So now we talk about visual analytics, meaning visual, so that can really tell you a story right away without much of interpretation. That's

1:00:55

* Carlo Lipizzi: kind of a a a media approach will never really happen, a a reality. But some visuals are more self explanatory than others. So you want to be sure that you you go in that direction. I remember

1:01:16

* Carlo Lipizzi: many years ago, a visualization of the traffic in Singapore.

1:01:30

* Carlo Lipizzi: So they came out with this visualization. We had the map of Singaporea changed the by the hour based on the traffic, meaning that the more traffic and the distance between Point a and Point B will be longer, less traffic a. B. We're closer.

1:01:35

* Carlo Lipizzi: So that's kind of a meaning stretching the the the actual map based on the the the traffic in in a particular moment.

1:01:54

* Carlo Lipizzi: That was particularly interesting, is an example very self-explanatory

1:02:07

* Carlo Lipizzi: visual, but with the embedded analytics. So there was the calculation of the traffic and mapping and the those metric in the the app

1:02:13

* Carlo Lipizzi: some examples. I will go super fast, whole part relationship. That's something that you can do. Discovery relationships, doing.

1:02:26

* Carlo Lipizzi: combined, explored, or in confirmation, analytics.

1:02:39

* merging together on multiple data types.

1:02:46

* Carlo Lipizzi: time, view of events, and analyzing the evolution in time. So all of those are a good examples of different types, all

1:02:52

* Carlo Lipizzi: visualization. So before we go into Python. I want to go for a moment to this book 202.

1:03:12

* Carlo Lipizzi: That is a a a quite interesting, a book, and I posted that

1:03:25

* Carlo Lipizzi: I will pause an extract. It's called the Out of Knowledge.

1:03:33

* Carlo Lipizzi: One of the things that I like on this book, a part of the visual that are standing is a

1:03:40

* Carlo Lipizzi: the attempt of creating a a classification for different visualization.

1:03:47

* Carlo Lipizzi: So there are quite a lot of different visualization.

1:03:54

* Carlo Lipizzi: But the certain point you need to define what are the user needs which kind of data is going to visualize? What are the the interactions that you may want to have.

1:03:58

* Carlo Lipizzi: So the order came out with this chat. There is one of the options. It's not saying that is the only one of the best, but it's definitely a good one

1:04:13

* Carlo Lipizzi: where you have a types and levels. So you have a statistical analysis, temporal analysis, and so on, and then the level. If it's a a micro mes or macro, based on that, that you have a different visualization that can serve you better.

1:04:23

* Carlo Lipizzi: I will add also.

1:04:42

* Carlo Lipizzi: there's a I mean I was intend to be for a a lines, but you will see different visualizations based on the different. It's kind of a a.

1:04:46

* Carlo Lipizzi: a, a 3, a flow chat that you can use to pick somehow the visualization, and it seems to be most appropriate for your particular case.

1:05:04

* Carlo Lipizzi: So those are some of the options again that the book is giving way much more than that. So

1:05:22

* Carlo Lipizzi: it's a it's a beautiful book, and then there are, I mean I for each one of those. Obviously there are a explanations that will give you all the details.

1:05:31

* Carlo Lipizzi: so visualization that somehow is an

1:05:43

* Carlo Lipizzi: So we have a a

1:05:47

* Carlo Lipizzi: a court. So that is, I am 622. That is a visualization and risk analysis

1:05:51

* Carlo Lipizzi: there.

1:06:01

* Carlo Lipizzi: Good!

1:06:03

* Carlo Lipizzi: It's a good course. It's been created by one of our engineering management professors. and I mean I'm. Mentioning that just to say that there is an entire a semester just on that visualization.

1:06:03

* Carlo Lipizzi: So going back to Python. So there are different packages that you can use. So Matt Plot Li, but is what we already know is a very basic visualization. Package is still the bread and butter for any visualization that you may want to have, and some of the other, including Cibona and in part bouquet, are based on Matt Plot lab.

1:06:21

* Carlo Lipizzi: It's not great. You cannot do very sophisticated things. You can now generate right away HTML page, but it's doing it's job.

1:06:47

* Carlo Lipizzi: See, Borne is more he is is a layer on top of that, but he's also using Pandas, and that means he's using the noon by meaning that you have an environment that there is more structure

1:06:57

* Carlo Lipizzi: Gg: plot that was originally developed for Oura. There is a python version no one is really using match but it it. It is there.

1:07:12

* Carlo Lipizzi: bookie. It's more recent, and the main advantage of book is that it's generating automatically the HTML.

1:07:24

* Carlo Lipizzi: If you want to do things that are more complex, like a dashboard blockly is probably the way to go. So let me go now in the next few minutes

1:07:33

* Carlo Lipizzi: on.

1:07:48

* Carlo Lipizzi: and you will have, by the way the solutions for the midterm in canvas in a moment.

1:07:52

* Carlo Lipizzi: So let me go into the visualization and let me start with the bouquet.

1:08:01

* Carlo Lipizzi: So that's how a bouquet.

1:08:10

* plain, basic code that I can be. So you import the the like, the libraries of sub libraries that you want to.

1:08:14

* Carlo Lipizzi: You create a file that will be generated.

1:08:22

* Carlo Lipizzi: You pass the characteristics of on the graph on the visualization you want to create. You pass the data. You specify what is the type of

1:08:26

* Carlo Lipizzi: visualization you want to create in this case is a line with those X and y value that's the with of the lane.

1:08:39

* Carlo Lipizzi: the line, and then you you I mean with this statement that you presented

1:08:47

* Carlo Lipizzi: on the screen. So if I run it

1:08:54

* Carlo Lipizzi: erez agmoni. So that's a good point, I mean is generating an HTML. So if you look at here, it's really an HTML page that means that you can export it. You can use it 150,

1:08:59

* Carlo Lipizzi: you you you can send to clients. You can embed it in into a website. Things like that

1:09:15

* Carlo Lipizzi: so it's kind of interesting.

1:09:24

* Carlo Lipizzi: So I have a bunch of samples, so so let me skip some of them. I will give you. Then you know the the the the script. So that's another one. Now let me run it.

1:09:27

* Carlo Lipizzi: So that's a little bit more complex. So you have multiple charts. So keep in mind that you can

1:09:43

* Carlo Lipizzi: zoom as you like you can reset it. You can download it if you want, but it is already in the HTML format.

1:09:54

* Carlo Lipizzi: So, just to make it more complex, so let me jump to the most complex. So that is this one. So you are giving a a more complex that is, the low as carve a more complex

1:10:04

* Carlo Lipizzi: points to dot on your screen that a. And you are using the bouquet for the visualization.

1:10:19

* Carlo Lipizzi: So what you will get will be something more complex like this one. So you can define again the caller. So the shape and obviously the the values in your visualization.

1:10:28

* Carlo Lipizzi: So those are all examples of. Let me now go to examples of seaboard. So see born a

1:10:43

* Carlo Lipizzi: a again that it

1:10:59

* Carlo Lipizzi: it's not as a sophisticated as a for something I mean in terms of visualization, not as sophisticated as as bouquet, because it's not generating any HTML,

1:11:03

* Carlo Lipizzi: but it's more sophisticated in terms of the underlying. Layer. It is because it is a Pandas.

1:11:18

* Carlo Lipizzi: It is known by. So in this case is basically working on a a data set that is relatively large. So let me run it up.

1:11:27

* Carlo Lipizzi: Thanks to

1:11:39

* Carlo Lipizzi: my powerful Mac is going fast. I'm kidding, so it's

1:11:40

* Carlo Lipizzi: pretty much what you would get with Matt Plotley, but but it's kind of nicer.

1:11:47

* Carlo Lipizzi: So finally.

1:11:54

* Carlo Lipizzi: let me talk for a second on the about plot, Lee. So there are 2 versions of plotley. One is open source; one is commercial, the commercial as a more feature.

1:11:56

* Carlo Lipizzi: but the open source is working just fine.

1:12:10

* Carlo Lipizzi: I mean, when you use it for commercial reasons, you may want to pay the commercial fees.

1:12:17

* Carlo Lipizzi: So in this case I'm using a data set from let me just run it just to see the differences.

1:12:23

* Carlo Lipizzi: So in this case, i'm analyzing schools. So if you think for a second.

1:12:33

* Carlo Lipizzi: I mean it's a relatively complex script. You you will have that, and is a analyzing universities by different accounts.

1:12:41

* Carlo Lipizzi: So, and it's generating the HTML so, and it's generating a in a kind of a cool way. So you have a multiple. So this is a a

1:12:52

* Carlo Lipizzi: pie chart. You have the buyer chart with the hoovering that will tell you

1:13:02

* Carlo Lipizzi: where they are from.

1:13:09

* Carlo Lipizzi: and then you have a line chart that is pretty much the same as the other. So there's what rank on the top 100. The universities, in terms of sanitation in terms of teaching.

1:13:13

* Carlo Lipizzi: so we are not there. By the way.

1:13:27

* Carlo Lipizzi: we would go there. So another example of Plotley

1:13:42

* Carlo Lipizzi: is a type of visualization that I like a lot. Because how about how much stories you can extract out of. That is a San key diagram

1:13:48

* Carlo Lipizzi: that was originally created to analyze the the the the flows, all that.

1:14:01

* Carlo Lipizzi: So it's easier if I run it, and you will say it.

1:14:11

* Carlo Lipizzi: It's what you have. So you have the starting point and the endpoint. So think about the traffic.

1:14:18

* Carlo Lipizzi: Think about the pipes. Think about concepts evolution.

1:14:25

* Carlo Lipizzi: So you can see how powerful this can be.

1:14:33

* Carlo Lipizzi: So we created something similar to understand that now how people generated that their knowledge in a given field. So from the very beginning to the user of that knowledge.

1:14:38

* Carlo Lipizzi: so it it's very cool.

1:14:51

* Carlo Lipizzi: You will have that in you are combust. So let me stop sharing now and let me go

1:14:55

* Carlo Lipizzi: to the in class assignment for 10 min, 1015 min, and then we we go back

1:15:08

* Carlo Lipizzi: they have me close, all of this.

1:15:18

* Carlo Lipizzi: I go here and in class assignment up.

1:15:24

* Carlo Lipizzi: Okay, and let me share this screen again.

1:15:30

* Carlo Lipizzi: All right. So then, in class assignment, is about working with the a. Csv file. That is, a collection of all the

1:15:38

* Carlo Lipizzi: tweets. I collected it

1:15:49

* Carlo Lipizzi: 2 years ago from the whole book, and you open the file. So the tweets

1:15:55

* Carlo Lipizzi: and the files being the file with the tweets and the file with the supports, You know, when that means with the so called the file into a list that that tweets into a Pandas structure remove the software. So using the software list

1:16:02

* Carlo Lipizzi: perform additional cleaning as needed, and then calculate and print the the top 10 words and the top 10 standards.

1:16:19

* Carlo Lipizzi: So let me

1:16:29

* Carlo Lipizzi: stop sharing. Let me post it.

1:16:32

* Carlo Lipizzi: Thank you.

1:16:44

* Carlo Lipizzi: Okay. I'm publishing everything.

1:16:50

* Carlo Lipizzi: all right. So i'm creating a

1:16:55

* Carlo Lipizzi: breakout rooms.

1:17:01

* Carlo Lipizzi: We have a 3 rooms, 2, 3 participants each. You have about 10 min 15 min to work on it.

1:17:06

* Carlo Lipizzi: The rooms are open now. See you in a bit.

1:17:15

* Carlo Lipizzi: Okay. So we are coming back. Sorry I didn't pose the recording, so if you will listen to the recording, you would see quite a lot of mapping.

1:30:03

* Carlo Lipizzi: So this about 1015 min of nothing, my apologies so another

1:30:15

* Carlo Lipizzi: 25 s, so all the rooms will be closed. So give us another few seconds, and then we will set

1:30:22

* Carlo Lipizzi: 5 s. Now.

1:30:43

* Carlo Lipizzi: Okay, so you are all back. You want to come back. Is there anyone who want to say something, present

1:30:50

* Carlo Lipizzi: the results, or just

1:31:00

* Carlo Lipizzi: sharing the experience?

1:31:03

* Carlo Lipizzi: All right, so let me share my screen. And this

1:31:09

* Carlo Lipizzi: Okay, so that's basically

1:31:19

* Carlo Lipizzi: what we have.

1:31:22

* Carlo Lipizzi: I imported pandas and a counter. I'm in the Probably this is a script that they did the

1:31:28

* Carlo Lipizzi: No. Totally recently. Probably, if I would do now, I would have done in a different way

1:31:39

* Carlo Lipizzi: erez agmoni. But that's the way it is. So reading into a Pandas data structure, the Tweets reading the soapore file into a lists 150,

1:31:46

* Carlo Lipizzi: then creating a

1:31:57

* Carlo Lipizzi: another data structure with the the send that the screen name from this one, creating a list

1:32:01

* Carlo Lipizzi: getting the top 10.

1:32:10

* Carlo Lipizzi: I mean that there are other ways that again, I would have done it in a different way. But that's the totally legitimate that is, is working fine.

1:32:13

* Carlo Lipizzi: So the top words I initialize a at least of a an empty list. Then I started

1:32:22

* Carlo Lipizzi: the iteration within the data structure.

1:32:35

* Carlo Lipizzi: a little bit of transformation or lower or alphabetical.

1:32:41

* Carlo Lipizzi: Then I created a a string out of the list, because a counter is taking a list, not a a string. And on the list.

1:32:47

* Carlo Lipizzi: Then i'm getting the the 50 most call mona, and that's going to be it

1:32:58

* Carlo Lipizzi: all right. So you have a the top 10 centers. So the top 50 words, and that's basically it. So questions

1:33:09

* Carlo Lipizzi: All right. So let me go into

1:33:26

* Carlo Lipizzi: in the assignment.

1:33:32

* Carlo Lipizzi: So the assignment is going to be pretty much on working on Pandas. So this assignment

1:33:43

* Carlo Lipizzi: is one of the not so many assignments, so that I didn't change in very recent times, because I really want to be sure, the students will know how to work with Pandas.

1:33:53

* Carlo Lipizzi: So you basically have a 3 types

1:34:07

* Carlo Lipizzi: to 3 sources of data. So movies users a rating. So it is all about movies. So the website is a a group L. So Dot or G, with all the information about the movies.

1:34:11

* Carlo Lipizzi: And this is not the most reason, the one that you have a a. As a file. So. but you have it. So there are some additional files to get some information about it.

1:34:28

* Carlo Lipizzi: But you basically have a movies.

1:34:45

* Carlo Lipizzi: users rating.

1:34:48

* Carlo Lipizzi: You have a occupation

1:34:51

* Carlo Lipizzi: for the different users. You want to

1:34:55

* Carlo Lipizzi: replace the numbers

1:34:58

* Carlo Lipizzi: from 0 to 20 with the what they are.

1:35:02

* Carlo Lipizzi: You want to pre into the last 5 rows. You want to find the 3 occupations

1:35:06

* Carlo Lipizzi: giving the highest rating for movies in the data. and that's basically if you want to merge the data

1:35:11

* Carlo Lipizzi: using this form of managing. There are several others.

1:35:20

* Carlo Lipizzi: You will use the txt fine that you have this moving Answer with me, Don't Txt, that will really drive you into how to do it.

1:35:26

* Carlo Lipizzi: Oh, again, that is, it is not

1:35:38

* Carlo Lipizzi: one of our most complex assignment, but is a good way to practice with pandas that are so essential for pretty much everything we do.

1:35:42

* Carlo Lipizzi: And I was not sharing. Oh, that's better.

1:35:58

* Carlo Lipizzi: Okay. help talking about that. So that's the homework.

1:36:01

* Carlo Lipizzi: So again, you have a

1:36:09

* Carlo Lipizzi: the Pdf. The Txc. That will give you additional information. You eventually have this website. If you want to know more. If you want to update the device and using the files that are more recent than those feel free to do that, so they are available in their website.

1:36:13

* Carlo Lipizzi: the structure. It's pretty much the same.

1:36:32

* Carlo Lipizzi: And that's basically what you what what you want to do. So you want to pre in to print the the 5 rows for each one of the other frames merge it.

1:36:36

* Carlo Lipizzi: print the number of records for each of the 4, the data frames.

1:36:46

* Carlo Lipizzi: and then you want to replace the occupation, and that it's originally in 0 to 20, with those

1:36:52

* Carlo Lipizzi: in the last 5 rows of the data frame.

1:37:00

* Carlo Lipizzi: and then up in the 3 occupation, giving the the higher the highest ratings for the movies on the the, the, the, the the the the

1:37:04

* Carlo Lipizzi: so that's basically it.

1:37:15

* Carlo Lipizzi: If you don't have questions. This the end of the class a little bit over time. My my apologies. It's 808, and

1:37:18

* Carlo Lipizzi: I will make sure that you have everything

1:37:30

* Carlo Lipizzi: again. We are still having a

1:37:33

* Carlo Lipizzi: here, and there are some issues with the population of the canvas shell that we are using for this course. So if you have any issue like the previous one, and the 2 answers in the quiz that we are pretty much identical. Send me an email, and I will act as as soon as possible, and definitely will have no impact on your grading.

1:37:38

* Carlo Lipizzi: If you see that it's something missing and let us know, and I will definitely address it right away.

1:38:01

* Carlo Lipizzi: Okay. So if you have any complaints, send me an email just to be sure. Send also an email to she you. So you have a a double possibilities to have your issue result.

1:38:09

* Carlo Lipizzi: Okay. So thank you all.

1:38:25

* Carlo Lipizzi: I really appreciate you being here, and we talk next week.

1:38:28

* Leona Chia: Thanks.

1:38:36

* Carlo Lipizzi: Thank you.

1:38:37