   
Hello. Hello, everybody. On the weather less.

0:43

 Hoboken less. So we had.

0:48

 A broken main yesterday.

0:55

 So it's more than 24 hours without water.

1:00

 So it's a lot of class. All right.

1:04

 So let's wait another minute or so, less than a minute, and then with start.

1:15

 Okay. So we have a relatively busy schedule for today, so we have an assignment to review.

1:29

 That was not an easy one.

1:42

 I would say that there were several things that we are not exactly mainstream compared to the previous assignments at least.

1:47

 And then we will talk about the.

2:01

 So quit engineering in a broad sense. We will talk about the visual intelligence and we will do an in-class exercise.

2:07

 One of you asked me to give more time for the in-class exercise just to have more options to practice.

2:20

 So I will do my best to be as fast as possible in the initial part, meaning the lights and the lecture and all the rest.

2:30

 So. Let me start sharing this screen.

2:44

 Welcome, everybody. So I'm sharing the screen.

3:00

 I would start close this.

3:05

 I would start with this.

3:10

 So in the previous assignment, there was a.

3:16

 Now. It's the trauma of the previous assignment that was on.

3:23

 The Sidbi files. So the main problem with those files was that we had two different files with two different formats.

3:32

 So one had the duration to the other, one had not yet all.

3:51

 They have us started and ended up the time.

3:57

 And the other difference was the.

4:03

 Use a type that in the old one was the name of the field user type and was a subscriber customer while in the new one.

4:09

 Instead of that user type member, underscore casual and you have a member or a casual.

4:23

 The other issue is that there were some characters that are a blank and I mean that.

4:30

 When you work with the real data, things can be messy.

4:43

 So when you work in a protected environment, all the records are true.

4:51

 But there are no missing values unless there's an exit size or missing values.

4:57

 And we think that there are no strange characters in in the content.

5:04

 I mean, in this case, in the cells and everything is fine.

5:13

 When you work with real world data, you get what you get.

5:19

 So if you work with social media, for example, when people is posting something, they character, etc., that they may use may be different.

5:25

 So if they are in one geographical region, you have a given character, etc.

5:36

 Another geographical reason is that region is a different character, etc.

5:44

 When you put them together and you use the US character, etc., some of those values,

5:49

 some of those characters may be recognized for something that you would not want to consider as such like.

5:57

 And then the line and the fine.

6:05

 So those characters that need to be cleaned somehow.

6:08

 So those are some of the issues that you may have when you deal with real world data.

6:14

 So in this case, there were some missing values that somebody else that we are not numerical

6:23

 know when they were supposed to be numerical or some other values that we are.

6:28

 Data, but may have not been recognized as a date with when you read into pandas.

6:36

 So those were the major challenges.

6:46

 So let me go here. I develop the solutions in two different ways.

6:50

 One with Lisa and one with pandas.

7:02

 So we introduce pandas last class.

7:07

 And some of you may be not that familiar with that approach and may want to use lists.

7:12

 So with the lists, things are more kind of convoluted.

7:23

 So just as a quick comparison, so there's the least version, any is a 129.

7:31

 I mean, obviously there are comments that are blanks, but all in all is under and 29 the pandas version.

7:38

 Same thing with all the comments and all the rest is at 79.

7:46

 So 79 compared to 129.

7:51

 So is a big difference. So.

7:57

 And then, unfortunately, the way the lists are working, when you work with lists, be prepared to do loops going into the content.

8:02

 When you work with pandas, you may not need the loops because you may have ways to do the calculation on the entire data frame.

8:15

 So with the lists, the idea is to read them those files into two lists of lists and then deal with whatever the content is going to be.

8:29

 Because we have been asked to use a function and to be named the present details.

8:47

 We want to be sure that they input that this function is going to get there will be the same.

8:58

 Even if the two files have a different structure and different content in terms of.

9:05

 I mean, how to name specific things like member casual or a user type.

9:12

 Meaning that before passing the values to the function print details if we need to create one single format for both of them.

9:20

 The print details. Function.

9:33

 It's relatively straightforward. So I initialize the all the counters that I will need.

9:39

 I started a loop for into the content.

9:50

 That is a list of lists with all the records that I read from the files with some changes that I will show in a minute.

9:56

 So you do you start the list, you add the duration so that this function is taking a duration.

10:11

 So just want to draw your attention to that is not end and start and end of the trip, but duration.

10:24

 So meaning that I calculated the duration before passing the value to the function start sation and station and that

10:33

 member or casual meaning I already changed from a customer and a user whatever was the name into member and casual.

10:44

 So. I increase the duration by the value in duration for each line.

11:02

 Then if is member one to the counter of members.

11:14

 If casual one to the counter of casual.

11:21

 Otherwise. Meaning if he's a blank or something like that, continue.

11:27

 Meaning doing nothing. So it is not meant, but not casual meaning.

11:30

 You have a blank. And we know that there are some blanks on that column to continue.

11:35

 Meaning doing nothing then. I'm checking if the name of the station where the tweet started is already in the dictionary.

11:43

 If yes, I will add one to the account. But we did something similar when we calculated the occurrence.

12:03

 The frequency of names in one of the in-class exercises is the same approach.

12:11

 I use a dictionary to count them. I mean, in pandas, you don't do that.

12:19

 You just count using a function embedded in pandas.

12:25

 That is a nonparty function. So.

12:31

 I mean, over here. I knew just that we already did a few times.

12:36

 I do not want to go back to the same then same thing for the end of the start, station and station.

12:40

 Same thing when I finished the lupus.

12:51

 I have all the information for that file and I can print that the average daily duration is and all the rest.

12:54

 Now I define a structure for the date that is a year to month date.

13:03

 Now what? Minutes and seconds. So you need this case.

13:10

 You need to define the format for the date.

13:15

 Opening the first file. Initializing a list that will contain the records will be, again, a list of police.

13:19

 Then I started the loop. If subscriber or customer.

13:31

 Start time is going to be. I will take the format that I define here to create a value with the start time.

13:44

 Same thing for the end time. Calculate the difference and then I add the to list the time difference and the value for the.

13:55

 For the state, stopping an interstate and ending station.

14:13

 Else continue meaning there is a blank or something.

14:22

 Uh, then, um. I'm asking you if subscriber.

14:29

 So this is basically to eliminate the basis.

14:36

 The blanks in subscribe for customer can be done in a different way.

14:41

 If subscriber, I will append. Member So I'm doing the transformation of transcriber to member.

14:47

 With this statement, that same thing, if casual would be is customary, would be casual, and then I will append at the end.

14:54

 I will have a list of lists with all the records within the first file.

15:04

 Pretty much the same with the second file,

15:12

 but the single file is going to be easier because I don't have the this portion replacing the subscriber with member and the customer with casual.

15:15

 I already has those elements, so that's why.

15:27

 Is it shorter? Then I call the function for the first one, call the function for the second one clause in the file not required.

15:31

 But I did it. And then I am printing and processing.

15:44

 So let me run it.

15:51

 So that's basically what you have a process in the first file average duration most populous start inflation and inflation casual users members.

16:00

 And then same thing for the second file.

16:12

 Again. That's with the lists.

16:18

 Without pandas, things would be faster with the pandas shorter.

16:22

 But there are some issues in particular.

16:31

 Keep in mind that if you are using Max and if you have one of the latest generation of a max or the one with the Apple silicon microprocessor,

16:35

 you may get errors. So I actually created a new directory.

16:50

 You see, that's one I'm in the. This one is the old one and they add to the warnings because otherwise I would get notifications of errors that

16:59

 were due to the non compatibility of that version of the version of Python that I was using on my old Mac before,

17:15

 but with an intel processor on the new Mac with an Apple processor and it didn't work well in some cases blocking the execution of the program.

17:24

 So I strongly encourage you, if you have this type of errors, go to the site on.

17:40

 Website and download the version of Python that is not the internal one but the other one.

17:48

 I don't remember how it's called, but you will check it if you have problems.

17:59

 Obviously just send me an email and I would send you the other way to go.

18:03

 So with pandas again it will be shorter.

18:10

 If you look at the function. So the function is basically just doing a small calculation on the time and then printing.

18:17

 So this case is taking a is taking a.

18:31

 Taken of over a generation. Most popular staff station.

18:40

 The most popular RN Station. A number of casual users and number of members.

18:44

 So opening the files and reading the science into pandas data structure.

18:51

 Initializing a list for the values for printing them.

19:02

 The one that I will pass to the PM dictates function.

19:10

 So what I'm doing here, I'm calculating it again.

19:16

 That's not the only way to do it.

19:22

 When I tried to do it on the all data structure, I got an error that most likely was related to the wrong version of Python.

19:25

 But I decided to change it and work on separate data frames for the different parts that they would need so they'd start.

19:40

 I calculated I transformed the value using the pandas to the time, specifying the format.

19:55

 Again, I could have done a different way using a just like I did for the previous one and variable with the format instead of rewriting it.

20:12

 So I did just two times. Doesn't change much.

20:21

 Same thing at the end. They put one on the stop start time.

20:24

 Stop time. Duration. It's just a difference.

20:30

 I do. You need to use this type because otherwise you will get an error or I mean, it is not something where you can do a calculation upon.

20:38

 And then I added the duration to that data for the search list that I will pass to the.

20:53

 A function that most frequent station.

21:03

 Super easy. So just from the dataframe that I created, I calculated the name that is according the most one line instead of a.

21:10

 It's very powerful. And then I added this value to the list that I will search for, for printing, same thing and station user type.

21:27

 Well, in this case, I don't need to do the conversion.

21:41

 So for the first one, I can calculate the number of members as a user type.

21:47

 So as I mean from the the call on user type and calculating subscriber.

21:54

 Meaning I no really need to do any conversion I calculating as members the records

22:02

 with the subscriber in the same thing customer and then passing it to the list.

22:11

 Similar thing for the second file. So same thing.

22:20

 On the first one, I'm not even considering the field duration.

22:26

 I could have done in a different way, but just wanted to have the same approach for the two files.

22:32

 So. Duration. Adding to the list the most frequent same thing that I did for the first one in this case cases number of members

22:40

 is a name taking a different name for what call on it because in the second one I have a different for the name.

22:51

 So the first one was user type, the second one is a member casual instead of looking for a subscriber and customer, in this case, a member and casual.

23:00

 So that's basically it. At this point.

23:13

 I have in data force all the elements on the first file, not the actual elements all defined, but all the values that they want to print.

23:16

 So it's a slightly different approach from the first one.

23:29

 The first one I was passing the entire file. So this way is more efficient and will generate the same results.

23:34

 Then calling the function with the fourth with the second that I could have done in a different way.

23:44

 Obviously, instead of having two separate. I could have a list of lists and doing a loop here on the two elements of the list.

23:52

 But I mean, it is just two elements and so doesn't make much sense and not saving that much space with it.

24:01

 So those was where the results from the non pandas version, the list version.

24:13

 This is going to be for. The new one.

24:22

 That is exactly the same. The only difference is that I have a different format for the duration because I wanted to have.

24:26

 I wanted to have Howard's minutes and seconds just to have a nicer printing.

24:40

 But I mean, it's basically the same value in a different format.

24:51

 Okay. Questions? I have one.

24:59

 Well, yeah. I mean, I would be surprised if no one had questions.

25:06

 It was a complex. I have a real one and then a bit of a silly one.

25:12

 Know. Is there going to be grading on this one?

25:16

 Because this gave me this is a real run for my money and faster in terms of like difficulty jump from the last couple of homeworks.

25:19

 Well, yes, there is there will be grading, but it will be kind of generous grading, considering that is more complex.

25:29

 Okay. And then I had another question that was actually homework or is that one kind of in the works?

25:43

 It was definitely a bit more hefty, I guess, than the other ones in terms of speed, but it'll take to grade.

25:52

 But I was just wondering. It should be graded sooner.

25:56

 I mean, our T.A. is grading all the assignments.

26:04

 We also have a greater. But we have two sections.

26:11

 4 a.m. 624. Okay. We have this section that is relatively small.

26:16

 And then we have another section with the 55 C. So that is running on Wednesdays, meaning they may have a little bit of issues on their.

26:23

 Okay. Thank you so much. And in particular, just to be very honest, we had several cases on the previous assignment of cheating in the other class.

26:37

 And so that's why we are a little bit behind.

26:51

 Just to let you know. So it took me almost a day to go through the entire process.

26:57

 So we have a tool that has been developed by University of Stanford that is comparing.

27:07

 I mean, you provide the to the tool, all the assignments,

27:19

 and they do automatically a comparison 1 to 1 with all the combinations at the

27:25

 very end that you have the number that will tell you the level of similarity.

27:32

 We used that level of similarity as a sort of red flag to determine if there could be plagiarism or not.

27:38

 Once we have the red flag, then we go 1 to 1 in that we manually analyze if there was a plagiarism or not because I mean,

27:52

 that is an automatic to learn, may work or may not work.

28:05

 And at the very end that is our responsibility to do something.

28:09

 Then I have a spreadsheet for evaluation.

28:15

 So with the grade for the assignment of the number of people doing the same assignment.

28:19

 In this case, we had the C six classes.

28:30

 So, I mean, it's a B class. We have the largest cluster with four people.

28:35

 There's more. Laura There's more or less cluster with two people.

28:42

 So one of the parameters in my model is the number of people involved in the project.

28:46

 So you have the the grade that can be hundred or a can be not so good, can be 80, whatever the number of people in that particular case.

28:54

 Then the evaluation from the tool and then the if the like.

29:07

 In that case, if there are two parts of the assignment, what is the relative weight of the two?

29:14

 I can probably go and show it to you.

29:24

 That is probably easier. No, I don't.

29:27

 Okay. So and at the very end, different things.

29:32

 Okay. So let me share this. So that's what we have.

29:45

 We have in this case, there were two parts.

29:52

 One, that was the coda and the other one that was the interpretation.

29:55

 So you have the score for each one of the two parts, the number of students that are involved.

30:03

 The percentage of similarity from the tool to the revised two, because again,

30:09

 we do 1 to 1 then is calculating what is the penalty and then the actual score for each one.

30:14

 Then you have the relative weight. You have the average.

30:23

 Because if. The average Joe dissimilarity is above the similarity that the students have.

30:29

 Then there is no penalty. But if the similarity in their submissions is higher than this will kick in.

30:43

 And then you have this or grade. So as you can imagine, this is taking a little bit of time.

30:55

 So my apologies, but I really wanted to give you all the insights.

31:03

 We want to be fair. So and we want to be clear in the process that we are doing so.

31:10

 But you can be sure that within a couple of days you will be graded.

31:18

 I appreciate the inside scoop. Thank you. This year, I mean, things can happen when you have a class that is 55 people coming from different cultures.

31:27

 Sometimes they really think that sharing solutions is something that you are not supposed to do.

31:38

 So I'm generally very clear at the very beginning of the class.

31:47

 But I mean, some students may not even be so fortunate.

31:52

 That's how. Other questions. Yeah.

31:58

 So I got a quick question. Yeah. Good for the quiz for quiz five.

32:04

 I know it's for you know, after I did it, it didn't tell you, like,

32:08

 which ones you got, right, which one you got wrong, like it just gave you the score. Didn't show you that the solutions is the one.

32:12

 Yeah. I don't know why that I will check it.

32:19

 I mean, it's just a matter of the box that probably has not been checked.

32:23

 The one of the questions that they had was on how you call, uh, the library number.

32:28

 And was a you and is right. Who asked me that.

32:39

 Yeah. I had mentioned that in my email yesterday. Yeah, yeah, yeah, yeah, yeah.

32:45

 So, I mean when you call a function, there is not one single way to do it.

32:49

 In that case, there was that one way.

32:57

 That was great. That is one of the many ways to do it and all the other that were wrong.

33:01

 Keep in mind that with by on capital letters sends more letters.

33:09

 They are different. So if you have the import non pi with the capital m is wrong because there is no library.

33:14

 No by with the with the capital n.

33:22

 So. But again, I'm sorry if there was no published solution for that.

33:27

 I will definitely check the box and make sure that it would be checked.

33:36

 Thank you. Sure. Sorry about that. Other questions.

33:43

 All right. So. Yeah.

33:51

 Yes. But the for this assignment and Simon finds out we're able to resubmit.

33:55

 Yeah. Yeah. Be sure that you are not using the solutions that I presented.

34:05

 Okay. Because, I mean, it wouldn't be fair. Of course.

34:13

 All right. Thank you. Okay. All right.

34:18

 So let me share the screen.

34:22

 And let me go now to the.

34:29

 File for the lecture. So let's talk about data science and machine learning.

34:34

 We do offer a course that is a machine learning for systems and enterprises is a course so that I created a couple

34:46

 of years ago and it's running once a year in fall and as am6 24 as a prerequisite because it's python intensive.

34:56

 I mean, m 624 or equivalent. So you need to be proficient in Python that would take the courts and is not one of your electives.

35:12

 But if you are interested, we can find a way some of your colleagues in the same cohort.

35:24

 The structure program took it and they liked it.

35:35

 Um, I'm also writing a book on a suicidal impacts of a machine learning,

35:41

 and some of the things that they would present today are somehow related to the work that I'm doing.

35:47

 What is artificial intelligence? So, I mean, we can talk for for days, months that the computers do not have intelligence.

35:56

 So no computer, no system that is being created.

36:08

 It's intelligent in the way we normally use intelligence as an attribute.

36:14

 But the point is just that we don't really know how to define national intelligence if we cannot define national intelligence.

36:23

 So we cannot define artificial intelligence, the history.

36:32

 But we have a mean test to evaluate the intelligence.

36:38

 So we're talking to me not about that. So the history of artificial intelligence is a starting back several centuries ago,

36:45

 just because there is no formal definition in the given the period of time when people realized that was an intelligent behavior

36:55

 or a behavior by the machine that was similar to what a human was doing then that that was labeled as an artificial intelligence.

37:06

 Generally speaking, artificial intelligence is considered to start with Alan Turing.

37:20

 So in 1950, Alan Turing was a.

37:28

 Working during World War Two and decrypting encrypted messages from the Germans.

37:37

 And it was a brilliant mathematician.

37:46

 And he created the Turing test. He created this decryptor a machine that was considered, for example, an intelligent one.

37:49

 Then if you move forward in that, let's say in the second half, in the 20th century, in the 1900s,

38:02

 artificial intelligence kind of took a little bit of momentum in the late eighties.

38:15

 I was involved in artificial intelligence.

38:25

 I was in a R&D company within an information technology group, and I was working on what we are called the time expert systems,

38:28

 meaning systems based on rules that the rules that we are representing, the expertise of the experts in a given field.

38:39

 So that was in 1986 when they started working on that.

38:56

 Then we realized that no matter how many roots we can create, we will never really have the same level of expertise as a human being then.

39:01

 Obviously there were other conditions, so we didn't have enough computing power.

39:13

 So. We didn't even have a computer with a graphical user interface.

39:19

 So the power of the computer were limited.

39:28

 We didn't have a python with all the libraries, meaning all the algorithms.

39:33

 So we had to write themselves ourselves by hand from scratch.

39:38

 Meaning? I mean, it was a lot of effort.

39:49

 Now you call a library, you pass the parameter, and you are good to go.

39:53

 Not at the time. So the limitations of the hardware resources, the limitations of the.

39:59

 Talk to the interpreters or copilot that we had at the time, the limitations of the data we had.

40:16

 So now everything is digital at that time.

40:26

 Very little was digital, although it was a lot of analog meaning how you can do machine learning or a data analysis if you don't have the data.

40:32

 So anyway, so then because of those limitations, there was what was called the winter of a yeah.

40:46

 So for a good 15 years, almost 20 years, nothing happened.

40:56

 And then with more data, more power and better languages, we started again.

41:03

 So this is just to give you a little bit of. The context and perspective.

41:11

 So artificial intelligence is not something that was born a few years ago, but is something that is relatively old.

41:18

 I mentioned the Turing test, so the Turing test is working pretty much like that.

41:29

 You have a wall and you have on one side the human being on the other side.

41:34

 We don't know. The human being is placing question to whatever it is on the other side of the wall.

41:42

 If no matter what the questions are, the answers are such that the human being cannot define.

41:50

 If on the other side there is a, uh, a human or an artificial entity.

42:01

 Then they are artificial and it is an artificial entity.

42:08

 Then the artificial entity passed the Turing test.

42:12

 Um, you will know that we have since on the health of November.

42:17

 Chuck GP That is a. On all over the news any is considering artificial intelligence.

42:25

 So people is using it for a million different things.

42:35

 It is a very powerful tool, but it's pretty much the equivalent of Google with the layer of, let's say, a conversational manager.

42:40

 So what is doing is taking a mountain, an ocean of data,

43:02

 and then discovering patterns in the data and matching the questions that you have with the partners.

43:09

 And based on the proximity of your question with the past, answer is giving you an answer in a conversational way.

43:21

 So it's a sort of compile the Google search answer presented in a conversational way.

43:29

 Would it pass the Turing test? Well, I asked Chuck Gupta, would you pass the Turing test?

43:40

 And the answer was absolutely right.

43:49

 It depends on who is asking the questions.

43:53

 Because if I am asking questions that are pretty basic, I may be bold.

43:57

 So like that they are called the lines, the larger language models.

44:04

 So what to do? The NLM may pass it.

44:10

 But then if the question questioner, the human questioner is more advanced, it wouldn't pass.

44:14

 So you may already have grabbed that in different outlets.

44:25

 But if you search, there are interesting examples of failures of charter.

44:31

 Some are kind of ironic. Some are funny.

44:40

 So it's kind of interesting. Anyway, a what is a a generally speaking, a part of automation.

44:45

 Not everything that is automated has intelligent.

44:55

 But if. If. You have something with artificial intelligence is part of the broader approach to automation.

44:59

 Robots can be intelligent or not.

45:09

 If you have a robot piling up a bulldozer and doing all that, we cannot really define it as intelligent.

45:12

 Within artificial intelligence, you may have autonomy or not.

45:24

 So if you have a group that has nothing to do with autonomy.

45:28

 If you have a so-called self-driving car that's never been there, that not happening.

45:34

 As we know from recent news that is on the line of autonomy.

45:40

 Drilling down artificial intelligence, machine learning, data science.

45:46

 So those are three components that are kind of interconnected.

45:51

 Not all like the machine learning or better not all the artificial intelligence is machine learning.

45:56

 Generally speaking, there are two schools, one that is called symbolic, that is based on symbols.

46:04

 Symbols can be a taxonomy.

46:12

 So taxonomy is a structured approach to knowledge.

46:15

 Consider the taxonomy of the human kingdom of the human or the animal kingdom.

46:21

 So you have animals. You have either no mammals, no mammals, and then you have all the different animals.

46:29

 So that's an example of a taxonomy of a classification of a domain.

46:39

 Each domain has its own classification taxonomy, and you can use the taxonomy as a way to navigate into the semantic meaning,

46:46

 the meaning of the documents or the images or whatever elements with the informational content that you want to analyze.

47:00

 So that's an example that I mentioned before, the expert system with rules.

47:16

 That's another example of a symbolic approach. So that's one way to interpret artificial intelligence.

47:21

 If you go back in time, this approach is based on that.

47:31

 So in the representation of knowledge, there were two different approaches.

47:36

 One that is based on the concept.

47:42

 We start with a blank slate and then we learn a.

47:48

 With experience. So those are the empty seats.

47:55

 And this is what the machine learning is doing. This is what it is doing from the data.

47:59

 All the knowledge is there. Then there are algorithms that are optimized to get to discover patterns in the data

48:08

 and present the pattern and associate the most appropriate patterns to your request.

48:20

 Is this a representation of a knowledge?

48:28

 Yes. Is this a representation of knowledge that we use?

48:32

 No. Meaning you cannot really define intelligent a system that is based only on a given data because it is not able to analyze new data,

48:36

 is not able to grow. It is just doing a matching though it discovering pattern and doing a matching.

48:53

 There are many other things that I would be happy to discuss with you, but I don't want to drive the entire class on that.

49:05

 So one is the empiricist machine learning.

49:15

 The other one is based on a concept that knowledge is something that is embedded in our brain.

49:19

 So we have a pattern. So we have a predefined model that we were already born with it.

49:32

 So and that's the symbolic approach. So one,

49:41

 but I mean that those two schools of thought where I mean that they started 2000 years ago or the traditional

49:45

 artificial intelligence or symbolic and the more modern machine learning approach are mapping 1 to 1,

49:57

 to schools of representational knowledge that are 2000 years old, obviously.

50:06

 We can do machine learning today because we have a lot of data.

50:14

 You couldn't do machine learning. When I started working in Artificial Intelligence in 1986,

50:18

 because there were not enough data to do battery recognition or to do any inference on the data.

50:24

 Data science is working for both. So when you have data, you need to process the data.

50:34

 What we did, for example, in the last assignment there was a cleaning the data, processing the data present in the data.

50:43

 So that's the real data science.

50:52

 Not everything in data science as an intersection with machine learning or artificial intelligence, but when you do in particular machine learning,

50:56

 it's kind of difficult to do machine learning without using elements that are typically originally from data science.

51:09

 In data science, you have all the data preparation, you have all the algorithms that you normally use in machine learning.

51:22

 So the intersection is pretty tight. Again, one of the key elements is the availability of data.

51:31

 Without data, you cannot do any machine learning.

51:40

 So machine learning. A system that is based on machine learning is to make main components, one that is the algorithm, one that is the data.

51:44

 Algorithms with no data will generate a model of no user data with no model.

51:58

 So you can do something, but it's not going to be a machine learning system.

52:06

 So again, that machine learning system is as good as the combination of algorithm plus data.

52:12

 What is more critical is the data.

52:20

 Keep in mind that on a chart GPT, they use an algorithm that is a I mean several are good, but the core algorithm is what is called transform.

52:23

 That is an algorithm created by Google and.

52:40

 It was used this transformative approach as a base for what is called GPP.

52:49

 That is the first instance of challenged jeep.

52:56

 So it is based on GP2, GP2.

53:05

 I mean, all of those are based on neural networks that are a sort of representation of the human brain.

53:11

 In the brain we have synapses in the neural networks, in the artificial neural networks.

53:25

 So you have the weights that you provide in the different multiplications that you do in the representation in GPT.

53:32

 Again, the GP2 is used and is using a half a billion of those weights of those parameters.

53:48

 We are in that GPT 3.5 with GPT four.

53:58

 That is due shortly with I don't remember how many billion parameters, probably 2.4 billion parameters.

54:04

 To train those systems to train a child GPU that is based again in on GP2.

54:16

 The course was a $1.5 million and as much energy as the energy required to run 1200.

54:28

 That's the number that they had in mind. But it's more than 1000 anyway.

54:40

 Medium sized cities in the US.

54:44

 Where I'm in. The energy is poor running.

54:49

 The computer's doing the calculation. So that's basically the difference between a GP and the GP.

54:54

 So they can only see GPT. They then building by the way, will have a tiny fraction of the data,

55:09

 meaning is not going to be as smart as the GP, GPT, even if the algorithms will be the same.

55:18

 So I mean the core algorithm is 300 lines, so binary code.

55:29

 So it's that much. But then you have several billion words collected from all over the Web.

55:35

 So those are some of the directions.

55:48

 In a robotics is definitely big.

55:55

 There are several flavors robotics,

56:01

 robotics and autonomous robotics or robotics to enhance human characteristics that can be physical or can be mental.

56:04

 Natural language processing definitely one of that creativity.

56:17

 I don't know about that. So Gardner is one of the leaders in.

56:22

 Market analysis and technology is doing those hype cycles where her technologies,

56:31

 where you start with the innovation, then pretty much everybody is talking about it.

56:44

 Think about. So they will reach the peak of expectations.

56:50

 And then either the technology will die or will become mainstream.

56:58

 So that's basically the. Life cycle of technology.

57:05

 And you would see quite a lot of things that we were discussing.

57:12

 You would see autonomous vehicles, cognitive systems and natural language processing.

57:17

 And then you have different calls for when Gartner believes that doing things is going to happen.

57:24

 Obviously, when you have elements that are more than ten years, then who knows?

57:34

 I mean, in in technology, something that is so far that they need we don't know what is going to happen.

57:42

 So if you look at the bottom left, the artificial leg in general, intelligence,

57:55

 meaning something like a system passing the Turing test with the examiner that is quite knowledgeable.

58:03

 It's far, far away. So we say more that then we have said, but we have no idea when it's going to happen.

58:17

 Same thing for autonomous vehicles.

58:23

 We know that Tesla has recalled most of the vehicles with autonomous I mean, the self-driving component.

58:27

 I mean, there are several complexity issues on there.

58:40

 But that's another view. And I would keep on that.

58:45

 Just driving the attention to the chart on the right.

58:51

 The number of public companies mentioning AA and machine learning in their earnings calls.

58:55

 So we went from zero to a huge number.

59:07

 Pretty much every one. We want to be in the AA machine.

59:13

 So it's events. Each school is, of course is on that a machine learning because each school wants more students.

59:18

 And when you have machine learning, students would come, oh, is this something good for the schools?

59:30

 Yes, it is something good for students. I don't know what mean is generating more confusion than ever.

59:40

 But that's another. A set of charts.

59:46

 Look at the bottom. Right. So the sentiment on the articles referencing a yeah.

59:52

 So from that skeptical period we are going to the enthusiastic period.

1:00:02

 So a lot of people is thinking that a could do.

1:00:08

 I'm in. I don't believe in a controlling what you do.

1:00:14

 I do believe that there will be high expectations, probably a use for what he cannot do.

1:00:22

 Generating results that could be disappointing.

1:00:36

 Hopefully will not be harmful somehow.

1:00:41

 Amazon is using elements of machine learning in software development.

1:00:46

 Machine learning can be used both ways, can be used in helping develop and developing software that can write code for you.

1:00:54

 I'm in. Feel free to use it. I use it for a few things, but you really need to know how to code to get a real benefit.

1:01:08

 From what Chad duped he can give you.

1:01:20

 By the way, not taking clients, meaning you cannot say okay, redefined, created a program to redefine and do something based on what is in the field.

1:01:25

 You cannot do that because cannot read clients. So in a sense it can be helpful in that way.

1:01:36

 But also software development can be helpful to develop better machine learning systems.

1:01:45

 You may know DevOps as term, meaning all the set of rules and methodologies to better develop software.

1:01:56

 When you develop software is not just writing the code but is having the code running in the system.

1:02:12

 In the larger system, when your piece of code will do his part, it's pretty much the same thing with machine learning.

1:02:22

 You have some code that will run within an environment.

1:02:35

 So this set of placing your code, the real world is a managing the operations of the code and has the equivalent.

1:02:39

 So a lot is on DevOps for machine learning.

1:02:54

 System engineering. Same thing, keeping in mind that when you develop something,

1:03:03

 even if a brilliant piece of technology based on machine learning is still a portion or something that can be bigger.

1:03:11

 So you need to keep the system view of what you are doing with that human so that somehow as to be in the loop.

1:03:22

 Just a brief example when you when they develop the chart the GPT.

1:03:31

 They had not one algorithm but was a method.

1:03:39

 So you have it. The first step is collecting this ocean of data.

1:03:44

 The second is to have human beings going into these additional data and doing two things.

1:03:52

 One, eliminating content that is inappropriate.

1:03:59

 And unfortunately, in the Web, there is a lot of content that is inappropriate.

1:04:04

 So those humans are doing a job that is one of the worst jobs that a human can do.

1:04:09

 Meaning going into the inappropriate content and remove it in five days, a week or more, 8 hours a day or more, just reading inappropriate content.

1:04:18

 So it is not a great job. And the second thing that they are doing up with tagging the.

1:04:37

 The area is machine learning with the highest number of humans employed is on net tagging data,

1:04:46

 meaning you have her text images eventually sound a new target for what they are

1:04:56

 that those will be to train the system when the system will read something similar.

1:05:04

 So the humans are cleaning, tagging a portion of the data, not the entire mountain.

1:05:13

 Then the system will use this first training portion to train a larger portion, and you humans will give scores to that.

1:05:22

 And then you have another element that is based on an algorithm that is called regarding neural network.

1:05:39

 That is basically using as a score to the proximity to the evaluation of the humans to evaluate the additional data that you may have at that point.

1:05:49

 You have a relatively large set of data as a training subset, and then you use that to classify all the rest and target in the proper way.

1:06:11

 So that's how the pattern of creation has been done.

1:06:28

 Once you have that, then you can apply systems for classification as the transform that was mentioning before.

1:06:34

 So again, it seems to be you have an algorithm with data, you apply, you're good to go.

1:06:43

 No is a system and you need to design the system in the appropriate way.

1:06:47

 Some applications are so. Uh self-parking.

1:06:53

 Cruise control or speech recognition, banks monitoring fraud.

1:06:57

 Those are examples of either machine learning or a, um, military just to share with you.

1:07:02

 Uh, an example I was attending, um, I'm working with the D.O.D. through circulatory system Engineering Research Center.

1:07:15

 That we have at the School of Systems and Enterprises.

1:07:33

 And we had a meeting presenting the research that we, as principal investigators, did for the day of the.

1:07:36

 And there was an invited guests who was talking about the some examples that he collected in a book that is called The War None The Army of Man.

1:07:48

 And he was he was serving in Afghanistan.

1:08:05

 And at the certain point, he was a close with the ledger, with the platoon, where,

1:08:10

 I mean, he was part of a and the certain point that a girl approached them.

1:08:22

 It was a nice build.

1:08:33

 They offer her some food. She stayed for a bit.

1:08:38

 Then she left. And then from the village, they started attacking them.

1:08:42

 So the question he had was if instead of us,

1:08:48

 there was an artificial intelligence evaluating the risk that that girl could have been sent to spy on us and then decide to kill the girl.

1:08:55

 Well, an artificial intelligence that would have done that.

1:09:11

 We didn't. So that's an example of human judgment.

1:09:16

 So there are autonomous vehicles.

1:09:24

 So some of the drones are both aerial drones or land drones.

1:09:28

 They are fully automated, but there are quite a lot of questions.

1:09:38

 So in the US we tend to have more ethical questions, even in our warfare.

1:09:43

 Some other countries may not. So we know that there are unmanned vehicles that are on the territory just acting as elements, active elements of war.

1:09:56

 So I would stop here and I would briefly go into something completely different.

1:10:18

 I will go very fast on this. So I will leave you this.

1:10:29

 Lights. So, software development.

1:10:36

 We know how it was created. One of the main issues when you develop software is that the specs are being lost in translation.

1:10:40

 So you have what the client needs, what the contractor understood, what the other of the subcontractor.

1:10:55

 And then at the very end, the delivery may be not exactly either what the client wanted or what the client asked them.

1:11:07

 So between what the client is asking, what the client it will want, and what the contractor is perceiving as a need,

1:11:19

 and what the I mean, all the parts are moving parts and there is quite a lot of lost in translation.

1:11:28

 So when you develop software, there is a portion of that is on development, the actual development.

1:11:38

 Fortunately, there is in testing and a portion of it is in maintenance.

1:11:44

 The sooner you find the harasser, the lower the cost of fixing is going to be.

1:11:50

 So one of the elements that we generally consider when we develop software is the concept of the software lifecycle.

1:11:58

 So the software as a. A beginning and an end.

1:12:07

 Some software apps thought of zombies. They never die.

1:12:13

 So we still have a COBOL software that is running since a few years.

1:12:17

 And it is running. It is running in public administration is running in some banks, but generally speaking you have a lifecycle and the retirement.

1:12:25

 So not always but most of the time that I tribally as a standard for that.

1:12:35

 So one of the traditional ways of developing software is a waterfall.

1:12:45

 So the waterfall meaning that you start with the requirements based on the requirements you do the design based on the design do decoding.

1:12:52

 When you complete the coding, you build testing and then you release the software to.

1:13:02

 The client, the annual start, the maintenance.

1:13:09

 So unfortunately, this approach, even if is great, for defining the requirements as a lack of flexibility for changes,

1:13:11

 making it not very successful, actually only 16% in the peak.

1:13:23

 This approach was called successful.

1:13:30

 So you can add a little bit of flexibility with a little bit of prototyping and later parameterization, but the problem is deeper than that.

1:13:35

 So that's basically when software engineering started the.

1:13:48

 With the idea of applying elements of engineering to engineering,

1:13:53

 design and development to software and computer science is essential for software

1:14:00

 engineering that physics is essential for electrical or mechanical engineering.

1:14:11

 I generally describe software engineering as a system.

1:14:20

 Engineering applied to computer science is basically considering the complexity of the software development

1:14:26

 that it is not just developing the components but having the components working in the proper environment.

1:14:39

 The The Software Engineering Institute that is currently based within Carnegie Mellon University in Pittsburgh.

1:14:50

 I was a partner member of a software engineer initiative to develop the capability maturity model.

1:14:58

 That is a way to define levels, to measure where you are and each level and help you moving up in the levels.

1:15:09

 It is very rare that you have a software at level four.

1:15:24

 Very few, if not none either. Higher than that.

1:15:32

 So those are a definition of a level. But in order to move from level C, you need to measure everything.

1:15:38

 Meaning you need to have a sort of big brother, big sister meaning quiet.

1:15:48

 Although all that adds to capture a score, analyze the information when you have a very large projector,

1:15:54

 the cost that the overhead may be justified by the size of the project.

1:16:04

 When you have smaller projects, not so much Agile.

1:16:12

 Agile was introduced kind of to keep a methodology, but with more degrees of flexibility.

1:16:17

 The GI Manifesto. So that's the main characteristics.

1:16:29

 We do not offer courses on Agile in a broad sense, so we do offer courses on Agile for software development.

1:16:35

 And we talk about agile development or agile management in our project management course.

1:16:45

 Principles. So pretty much you interact a lot.

1:16:55

 You have a lot of prototyping. So those are some of the methods.

1:16:59

 Scrum is one of the most common. And.

1:17:06

 I would stop here because it's, uh, it's later than 745 and you have the lights and you can definitely go through them.

1:17:15

 I strongly encourage you also to go through there's lights and I mean, they are relatively old,

1:17:30

 but is a description of one of the largest failures in software development.

1:17:38

 The in this case was by the FBI.

1:17:44

 The ritual case file that court said asks that it be a huge amount of money generating not much.

1:17:49

 So spoiler alert the real reason is the lack of communication and lack of clear responsibilities.

1:18:02

 So again, it's not just a matter of how you develop software, but how you manage your bank.

1:18:12

 And so from the client to the developer, the environment, the real needs.

1:18:20

 So it's kind of interesting.

1:18:27

 I use the original slides that were published in 2005 but didn't change anything because I just wanted to give you.

1:18:31

 An example of what's going on. All right.

1:18:41

 So in class exercise. So the in class exercise, it's about analyzing a password.

1:18:45

 So the user will impose a set of passwords separated by a comma.

1:18:55

 And you need to check if all those conditions are verified, meaning the password is at least one letter between those,

1:19:04

 one number between those, some of the capital letters, some special characters, minimum length of maximum lengths.

1:19:14

 And you will take from the user the.

1:19:27

 Passwords. Most likely, at least.

1:19:34

 And then you look into the list. You would write a function of.

1:19:38

 That will take each one of the passwords and do the checks.

1:19:45

 And then returning is good. Is not good.

1:19:50

 All right, so I'm stopping the sharing that I'm publishing.

1:19:55

 The in-class exercise.

1:20:07

 SHUBERT Yes. Well. And.

1:20:18

 I will create the breakout rooms. Four breakout rooms to three participants, creating opening 15 minutes.

1:20:25

 All right. The only 2 minutes. Then 15 minutes.

1:20:41

 Hi. Sorry, my laptop died. Can you put me back in the room?

1:30:51

 One. Okay. Hold on a sec.

1:30:55

 I sprinted to get my charger so I can go back there.

1:31:04

 Okay. Thank you.

1:31:09

 Sure. Resuming the recording.

1:31:12

 Okay. So we are all back. Any one of you want to share either the solution or just the frustration of not having enough time to do it?

1:32:31

 All right. Okay, so let me.

1:32:55

 Share the screen and let me do.

1:33:01

 Here. So.

1:33:06

 I use the the library string.

1:33:13

 Uh, you can do without.

1:33:17

 So let's see both the options. So in the comment, I added all the requirements.

1:33:21

 Then I imported this string. You would see how this would be used.

1:33:29

 And then this is the function validating the password.

1:33:35

 So it's taking a single password then is checking if is either less than six characters or more than 12.

1:33:43

 If this is the case, it will return an.

1:33:54

 Dan is transforming the world into a place when you transform a stranger into a least the elements on the list,

1:34:02

 that will be the characters, the single characters.

1:34:13

 So one element, one character from the origins thing.

1:34:16

 Then I created a list that I will use for getting the okay for each one of the tests.

1:34:23

 So there are four tests to do as more letters, numbers, capital letters, special characters and order each one.

1:34:33

 I will eventually change the RN into White.

1:34:47

 So the first one I am checking if the element has a I mean, I need to go a little bit down.

1:34:52

 So I created a list for small letters, capital letters, member numbers.

1:35:04

 This can be done using the function three anchor or you can manually create.

1:35:13

 So in comments I use the manual creation on those lists.

1:35:19

 Out of the comments I use the function string that it was mentioning before.

1:35:24

 So those are the values. So. And checking if it's more letters.

1:35:30

 If a is a number is numbers, if capital letters, special characters, each one is an okay.

1:35:40

 Then we'll change that and into y and then.

1:35:53

 If it's and then we're done the list.

1:36:00

 I mean, if no one of those or those if we generate it.

1:36:06

 Yes. You will get no other otherwise.

1:36:14

 Yes. So when you go down.

1:36:18

 You pass the elements.

1:36:26

 So it's taking the password. Then there is a passenger, each one of their passwords into the function, and then if it is returning, why then is okay.

1:36:31

 Otherwise we'll continue. Who let me save the cops.

1:36:50

 This list here, and I will use it as an example.

1:37:01

 So let me run it. Back.

1:37:09

 And as a result. All right.

1:37:28

 So you will find the two more.

1:37:33

 Filed by those creeps. Just play with them.

1:37:39

 But they will get back the next class to explain a little bit more what they are.

1:37:45

 So they are basically to generate reports or to analyze the files for what is called data, exploratory analysis.

1:37:50

 So one is doing the full exploration.

1:38:03

 The other one is a focus on the coordination on.

1:38:07

 All right. So. Next assignment, the.

1:38:13

 It's going to be. On COVID 19.

1:38:19

 So the file that you will use for your analysis will be, What is it?

1:38:24

 Will be like this one. So you have.

1:38:35

 Condition groups specific condition.

1:38:41

 A code that you would not use the age group.

1:38:45

 The number of deaths and the number of mentions.

1:38:49

 So this is the file you have. So based on that, the.

1:38:53

 You will do.

1:38:59

 I mean, let's keep but one of is just questions.

1:39:04

 So you will basically calculate the number of people in the different categories.

1:39:09

 So where are categories it would be?

1:39:19

 So you will not read the file. You will create a dictionary, or you want to use pandas and feel free to do it.

1:39:25

 Population less than 35. Then you will.

1:39:36

 Plot the. For two elements.

1:39:44

 So one bar plot and a pie plot similar to this one.

1:39:51

 But this one is from a different example. So not the actual graph.

1:39:56

 And we'll show the distribution of deaths by age group, meaning you want to calculate.

1:40:03

 So create a dictionary with the number of deaths for each one of the categories.

1:40:10

 You would use a function. So.

1:40:17

 You will then generate the chart.

1:40:25

 Bar by. Ukraine, their co-morbidity with the highest number of deaths with a population of less than 35.

1:40:32

 Run correlation analysis to determine the relationship between variables.

1:40:43

 You will have your script.

1:40:48

 And then there will be part one. That would be just questions.

1:40:54

 So what is a giant software? What is a why?

1:40:59

 Waterfall is not a great solution to define machine learning.

1:41:03

 And some examples. So that's basically it.

1:41:09

 If you have time, adjust and eventually you will have the recording, considering you will use the correlation.

1:41:16

 I just want to be sure that you will see it running at least once.

1:41:25

 So correlation. In this case, I'm using a basic pandas functions and my plot labor to.

1:41:31

 I have the graphical representation of the keyboard and is another graphical library on top of my plot.

1:41:44

 So I'm reading the NFL Census CSB file.

1:41:54

 Calculating the correlation and representing it as a graph and eventually save it has a point.

1:42:01

 If I run it. The fine is relatively big.

1:42:11

 You have the correlation. So you will see that the salary, for example.

1:42:17

 So the correlation is a negative correlation, meaning one is decreased is increasing, the other is decreasing, and vice versa.

1:42:23

 Positive one is increasing and the other is decreasing. So the more blue is, the more positive.

1:42:32

 So obviously along the main diagonal there is a one.

1:42:39

 So you have the salary.

1:42:44

 That will increase with experience and if the Pro Bowler will be higher.

1:42:50

 That's an example. And you will have the file that would be generated.

1:43:00

 The other one is a little bit more powerful, even if a few lines.

1:43:11

 So. Is generating.

1:43:17

 A reporter. So it's creating it.

1:43:23

 Summarizing the dataset. Generating the report.

1:43:28

 Generating an edge male end is basically something that.

1:43:35

 Not sure where this one comes.

1:43:47

 Go here for a moment. But.

1:43:53

 So there's the excitement that is being generated.

1:44:11

 It's I mean, it's amazing that in Q line to line, basically you can generate the entire report with overview description of variables.

1:44:15

 Distributional numbers. Analysis for each one of the variables, including the distribution.

1:44:27

 To evaluate the skewness. Then you have.

1:44:39

 Interaction correlation pretty much as the other one.

1:44:50

 Either as a heat map or a stable. Missing values and a sample.

1:44:56

 So all of this. With pretty much one library.

1:45:05

 And one. I mean, there's three things. And one line.

1:45:14

 Two lines. And then if you want the five, three lines.

1:45:19

 Okay. So I will post all the content again.

1:45:25

 I already send an email to the DEA and the grader to make sure that the evaluation will be less strict as the previous one for text size for.

1:45:33

 And you will have another couple of days for your submission.

1:45:48

 Not that you need to do a late submission, but you have the option.