NYU Center for Data Science

DS-UA 112
Introduction to Data Science

Week 6: Lecture 2

Text – Patterns of Characters

NYU Center for Data Science

How can we process strings in text?

DS-UA 112
Introduction to Data Science

Week 6: Lecture 2
Text – Patterns of Characters

Adapted from Hug, Sedgewick, Wayne

Announcements

- Please check Week 6 agenda on NYU Classes
  - Homework 3
    - Question 3
  - Lab 6
- Remember to post to Piazza

See @240 on Piazza for more information about Question 3a

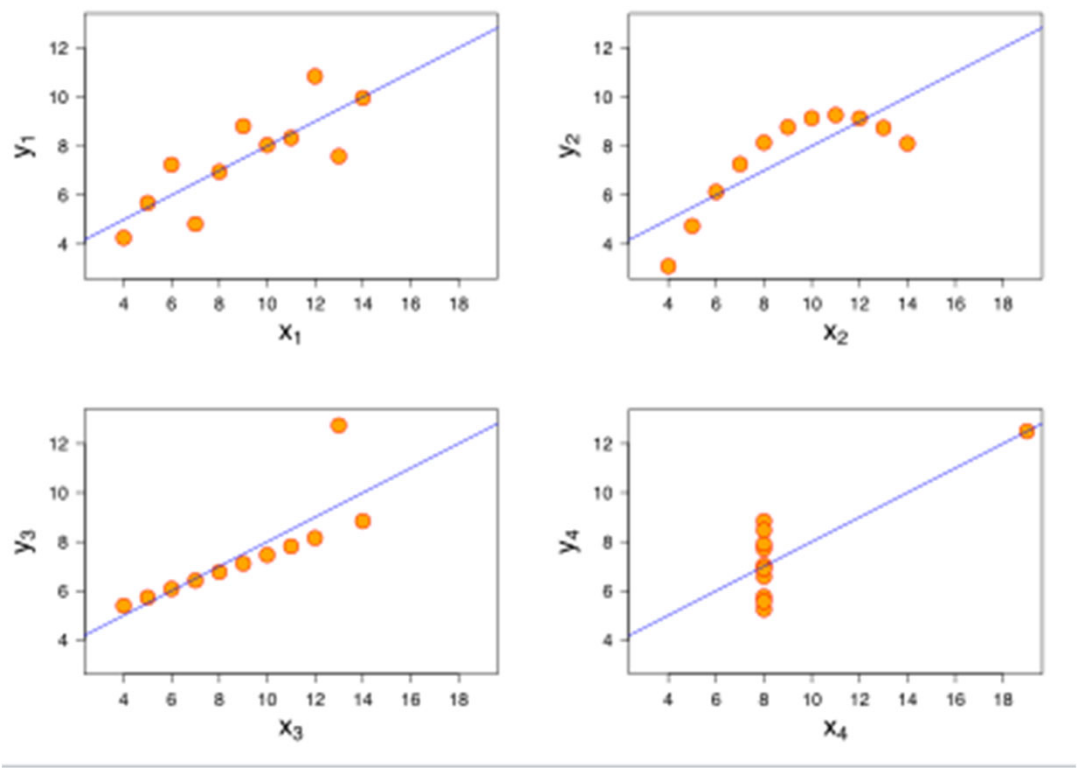
Announcements

- Midterm
  - Wednesday March 11
    4:55-6:10pm
  - Reference Sheet
- Practice
  - Exam
  - Problems
  - Examples

Bring your reference sheet along with our reference sheet

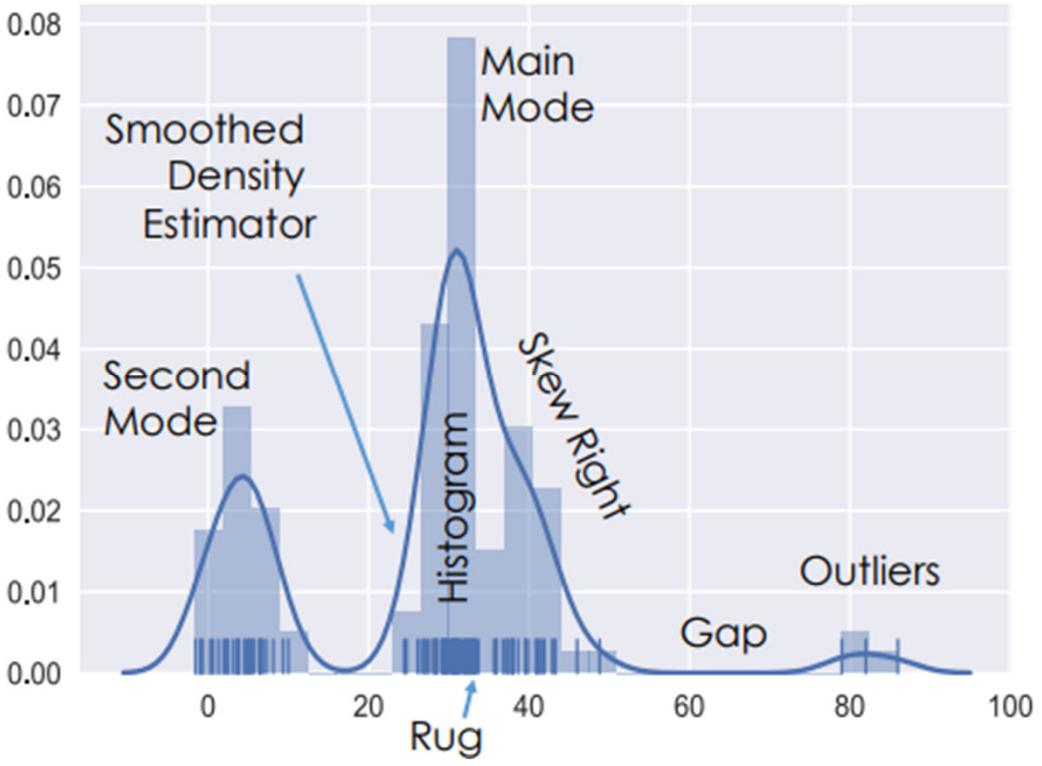
Review

- Types of Charts
  - Single Variable
  - Multiple Variables
- Approaches to Visualizations
  - Scale
  - Conditioning
  - Colors
  - Transformations



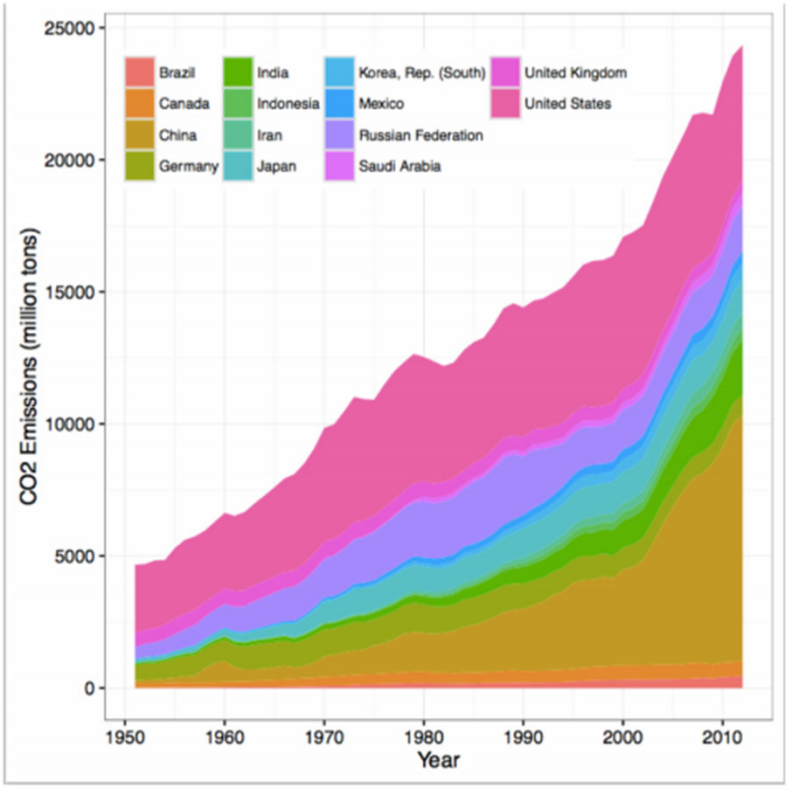
Review

- Types of Charts
  - Single Variable
  - Multiple Variables
- Approaches to Visualizations
  - Scale
  - Conditioning
  - Colors
  - Transformations



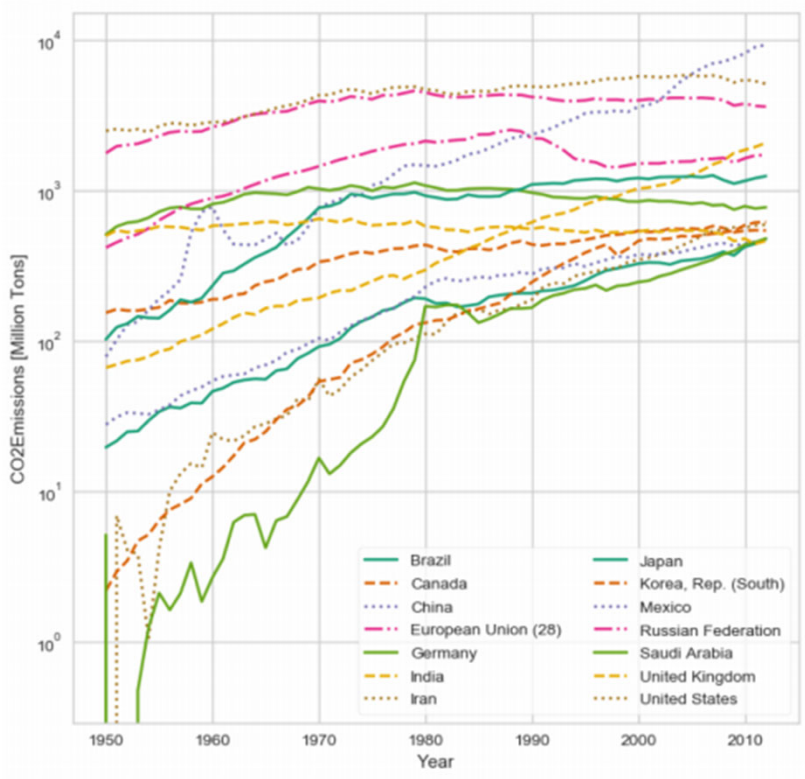
Review

- Types of Charts
  - Single Variable
  - Multiple Variables
- Approaches to Visualizations
  - Scale
  - Conditioning
  - Colors
  - Transformations



Review

- Types of Charts
  - Single Variable
  - Multiple Variables
- Approaches to Visualizations
  - Scale
  - Conditioning
  - Colors
  - Transformations



Agenda

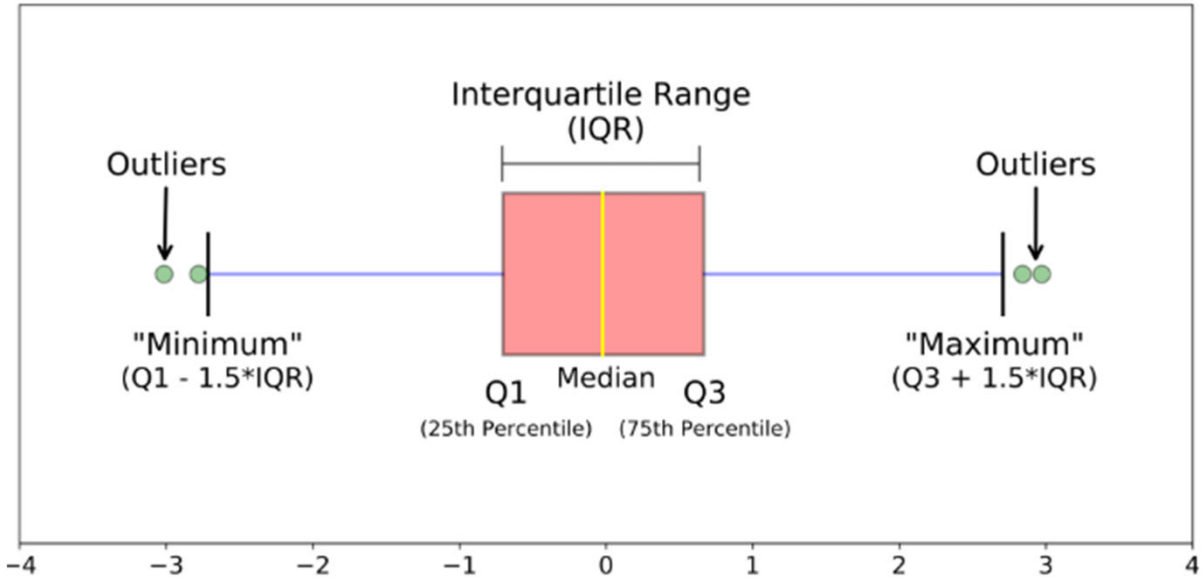
- Boxplots
  - Percentiles
- Operations with Strings
  - contains
  - split
  - extract
- Regular Expressions

References

- Nolan, Lau, Gonzalez (Chapter 8)

Boxplots

- Boxplots are charts used to describe a single quantitative variable
- Percentiles are subsets of the numbers in sorted order
- Quartiles are the
  - 25th percentile
  - 50th percentile
  - 75th percentile
  - 100th percentile



String Operations

- Remember the problem of finding names in New York state that start with the letter “E” or end with the letter “y”?

```
starts_with_j = [x[0] == 'J' for x in babynames["Name"]]
babynames[starts_with_j].sample(5)
```

Sex	Year	Name	Count
F	1945	Jannie	8
F	2005	Jolina	9
M	2009	Jaylan	8
F	1993	Joseph	26
M	2010	Jess	5

- We could determine the rows with a list comprehension giving True or False

String Operations

- Remember the problem of finding names in New York state that start with the letter “E” or end with the letter “y”?

```
babynames[babynames["Name"].str.startswith('J')]
```

Sex	Year	Name	Count
F	1945	Jannie	8
F	2005	Jolina	9
M	2009	Jaylan	8
F	1993	Joseph	26
M	2010	Jess	5

- We could determine the rows with a list comprehension giving True or False
- Another approach would be to use the str methods for Series

String Operations

- Series have many helpful operations on strings included in str methods

```
babynames[babynames["Name"].str.contains('ad')].sample(5)
```

Sex	Year	Name	Count
F	2018	Madelyn	336
F	2018	Guadalupe	98
M	1984	Bradford	32
F	2000	Khadija	5
F	1995	Soledad	31

- The method contains will return True or False depending on whether the string contains the characters

String Operations

- Series have many helpful operations on strings included in str methods

```
babynames["Name"].str.split('a').to_frame().head(5)
```

	Name
0	[M, ry]
1	[Helen]
2	[Dorothy]
3	[M, rg, ret]
4	[Fr, nces]

- The method split will return split the string into many strings depending on a character
- We can generate either a list with the string or additional columns for each string

String Operations

- Series have many helpful operations on strings included in str methods
- The extract method with identify the capturing group in a regular expression allowing us to pull out small strings like words from large strings like sentences
- Regular expressions allow use to match patterns leading to many applications
  - Access information in digital libraries
  - Filter text such as spam emails
  - Validate data-entry fields such as date or URL

See documentation for more information

Regular Expressions

operation	order	example	matches	does not match
concatenation	3	AABAAB	AABAAB	every other string
or	4	AA|BAAB	AA BAAB	every other string
closure (zero or more)	2	AB*A	AA ABBBBBBA	AB ABABA
parenthesis	1	A(A|B)AAB	AAAAB ABAAB	every other string
		(AB)*A	A ABABABABA	AA ABBA

Exercise

- Which of the following would match the with

"jo*hn"

Try it on regex101.com which provides a platform for experimenting with regular expressions

1. jooohn
2. jon
3. jhn
4. john
5. jooooooohnn

Regular Expressions

operation	example	matches	does not match
any character (except newline)	`.U.U.U.`	`CUMULUS` `JUGULUM`	`SUCCUBUS` `TUMULTUOUS`
character class	`[A-Za-z][a-z]*`	`word` `Capitalized`	`camelCase` `4illegal`
at least one	`jo+hn`	`john` `jooooooohn`	`jhn` `jjohn`
zero or one	`joh?n`	`jon` `john`	`any other string`
repeated exactly {a} times	`j[aeiou]{3}hn`	`jaoehn` `jooohn`	`jhn` `jaeiouhn`
repeated from a to b times: {a,b}	`j[ou]{1,2}hn`	`john` `juohn`	`jhn` `jooohn`

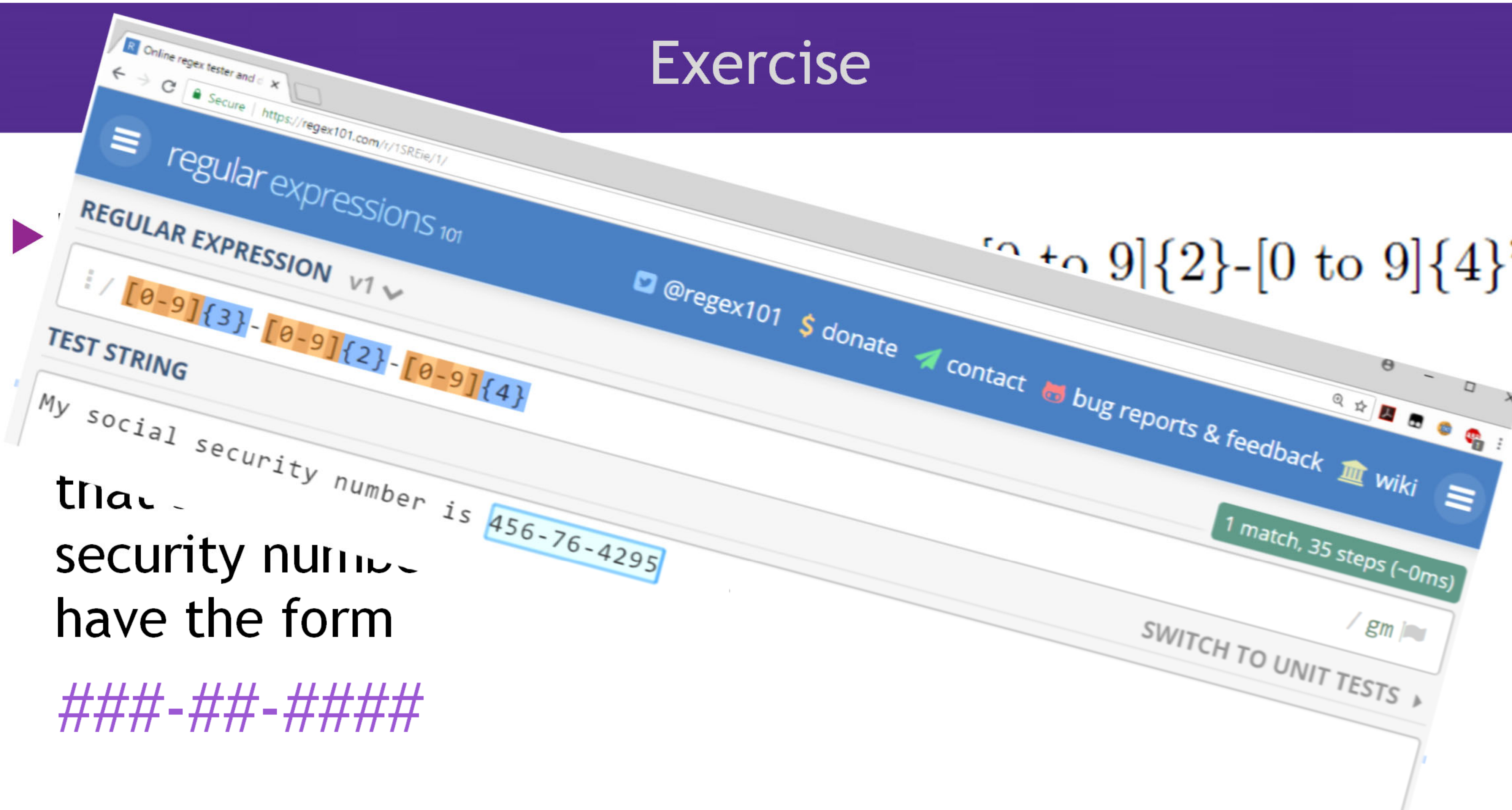
Exercise

- How could we match social security numbers in text. Remember that social security numbers have the form

  ###-##-####

1. "[0 to 9]{3}-[0 to 9]{2}-[0 to 9]{4}"
2. "{0-9}[3]-{0-9}[2]-{0-9}[4]"
3. "[0-9]{3}-[0-9]{2}-[0-9]{4}"
4. "[0-9]{3}-[0-9]{3}-[0-9]{3}"

Exercise



▶ … [0 to 9]{2}-[0 to 9]{4}”

that … security numb… have the form

###-##-####

Regular Expressions

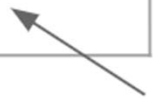
regex	matches	does not match
`.*SPB.*`	RASPBERRY CRISPBREAD	SUBSPACE SUBSPECIES
`[0-9]{3}-[0-9]{2}-[0-9]{4}`	231-41-5121 573-57-1821	231415121 57-3571821
`[a-z]+@([a-z]+\.)+(edu|com)`	horse@pizza.com horse@pizza.food.com	frank_99@yahoo.com hug@cs

Regular Expressions

operation	example	matches	does not match
built-in character classes	`\w+` `\d+`	`fawef` `231231`	`this person` `423 people`
character class negation	`[^a-z]+`	`PEPPERS3982` `17211!↑å`	`porch` `CLAmS`
escape character	`cow\.com`	`cow.com`	`cowscom`

Regular Expressions

operation	example	matches	does not match
beginning of line	`^ark`	`ark two` `ark o ark`	`dark`
end of line	`ark$`	`dark` `ark o ark`	`ark two`
lazy version of zero or more *?	`5.*?5`	`5005` `55`	`5005005`



5*5 would match this!

Questions

- Questions on Piazza?
  - Please provide your feedback along with questions
- Question for You!

How can we match patterns of characters in strings?

Regular Expression Tutorial
https://docs.python.org/2/howto/regex.html

Summary

- Boxplots
  - Percentiles
- Operations with Strings
  - contains
  - split
  - extract
- Regular Expressions

Goals

- Understand divisions of numbers into percentiles for boxplots
- Use string methods on tables to process text
- Find patterns of characters with regular expressions

Questions

- Que
  - [
- (

How can we mate
patterns of characters in
strings?

https://doc x.html