# DATA 100: Vitamin 4 Solutions

February 19, 2019

## 1 Kernel Density Estimation

A kernel density estimator plot can be used in lieu of which which of the following visualizations:

- ☑ A histogram
- ☑ A boxplot
- ☐ A scatter diagram
- ☐ A bar chart
- ☐ None of the above

**Explanation:** A KDE plot is used to visualize a continuous quantitative variable. It can therefore be used in lieu of a histogram or a boxplot, since both of these visualization methods are also used to plot a single continuous variable. A KDE plot can not replace a scatter diagram, since they are used to plot a minimum of two quantitative variables. A bar chart is used to visualize discrete or qualitative variables.

# 2 Regular Expressions

## 2.1 Matching

Which of the following strings will fully match this regular expression pattern: `^Ex[^aeiou]+\w*[ ]*\w*\??$`

- ☐ Exercise?
- ☐ Exercise
- ☐ Exer
- ☐ Ex
- ☑ Extra
- ☑ Extra fries?

**Explanation:** Try out each answer option at https://regex101.com/.

## 2.2 `re.findall`

Given the following string, `text = "My favourite numbers are 1, 14 and 120.`, which of the following lines of code will return the following result: `["1", "14", "120"]`.

- ☑ `re.findall(r"[0-4]+", text)`
- ☐ `re.findall(r"[1-9]+", text)`
- ☐ `re.findall(r"[0-9]", text)`
- ☐ `re.findall(r"[0-9]*", text)`
- ☑ `re.findall("\d+", text)`

**Explanation:** Try the code out for yourself!

# 3 SQL

## 3.1 How

For which text values `x` below would this `LIKE` expression be true?

```
x LIKE dog%
```

- ☑ dog named cats
- ☑ dogs and cats
- ☐ cats and dogs
- ☐ cat and dog
- ☑ dog

**Explanation:** `x LIKE dog%` finds any values that start with `dog`.

## 3.2 Why use databases?

Given a dataset, for which of the reasons listed below should we consider using a database instead of storing them in Pandas DataFrames?

- ☐ The dataset is hierarchically structured.
- ☐ The dataset is not in tidy format.
- ☑ The dataset is too large to fit in main memory.

**Explanation:** Databases are highly optimized for storing structured data efficiently. They should are also essential when the size of the data is too large to be stored Pandas DataFrames. Whether the dataset is hierarchically structured or in tidy format should not be afactor in your decision to use a database.