pandas_basic_data_cleaning_workbook

September 22, 2020

1 PANDAS BASIC DATA CLEANING WORKBOOK

Have a go at the following questions to practice your new found skills.

If you have any questions, go back to the course videos and have another look. One version of the answers is available in the next download. Remember, there are different ways to write code to get the same answer, so your answer can be correct and different to the answer example!

If you feel stuck and want some in person help, then have a look at the events page to join in a workshop https://swamphen.co.uk/events.

```
In [ ]: # data from
        # https://www.kaggle.com/aparnashastry/building-permit-applications-data
```

```
In [ ]: # import pandas and numpy
```

```
In [ ]: # read in data
```

```
In [ ]: # print out the head
```

```
In [ ]: # check the info
```

```
In [ ]: # check how many items in each column are null, i.e. a nan value
```

```
In [ ]: # why do you think there is so much missing data?
```

```
In [ ]: # what would be the best way to fill the missing data in the zipcode column?
```

```
In [ ]: # select the entries with Street Suffix equivalent to nan
```

```
In [ ]: # pull out the information for Street Name and Street Suffix for
        # the first nan identified above
```

```
In [ ]: # is the answer to the missing data in the street name?
```

```
In [ ]: # fill in this Street Suffix with the correct information from the street name
```

```
In [ ]: # check this has filled in the info ok
In [ ]: # you could do all the other missed entries like this
        # but there are over 2000 missing entries
        # instead, just fill the other missing Street Suffix with the Street Name
```

```
In [ ]: # check you have filled all the missing values
```

```
In [ ]: # what is the most frequently occuring permit type?
```

```
In [ ]: # find the missing permit types and fill them with  the most frequently
        #occuring permit type number
```

```
In [ ]: # how big is your data set?
```

```
In [ ]: # remove all the rows that contain a nan
```

```
In [ ]: # how big is your data set now?
```

```
In [ ]: # is this a suitable approach for this data set?
```