# pandas_basic_data_cleaning_workbook_answers

September 22, 2020

## 1 PANDAS BASIC DATA CLEANING WORKBOOK ANSWERS

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
        import pandas as pd
        import numpy as np
```

```
In [ ]: # read in data
        building = pd.read_csv('Building_Permits.csv')
```

```
In [ ]: # print out the head
        building.head()
```

```
In [ ]: # check the info
        building.info()
```

```
In [ ]: # check how many items in each column are null, i.e. a nan value
        pd.isnull(building).sum()
```

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

        # covers all sorts of different projects where the buildings may
        # not have that storie or suffix
```

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

        # look it up and fill it in
```

```
In [ ]: # select the entries with Street Suffix equivalent to nan
        np.where(building['Street Suffix'].isnull())
```

```
In [ ]: # pull out the information for Street Name and Street Suffix for the first nan identif
        print(building['Street Name'][143])
        print(building['Street Suffix'][143])
```

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

        # yes
```

```
In [ ]: # fill in this Street Suffix with the correct information from the street name
        building['Street Suffix'][143] = building['Street Name'][143][12:16]
```

```
In [ ]: # check this has filled in the info ok
        print(building['Street Suffix'][143])
```

```
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

        missing = np.where(building['Street Suffix'].isnull())
        print(missing)
```

```
In [ ]: for missed in missing:
            building['Street Suffix'][missed] = building['Street Name'][missed]
```

```
In [ ]: # check you have filled all the missing values
        np.where(building['Street Suffix'].isnull())
```

```
In [ ]: # what is the most frequently occuring permit type?
        building['Permit Type'].mode()
```

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

        # there are no missing permit types
        np.where(building['Permit Type'].isnull())
```

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

```
In [ ]: # remove all the rows that contain a nan
        building = building.dropna()
```

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

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

        # no!
```