eda_vijay

December 9, 2018

In [1]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
%matplotlib inline
```

In [2]:
```
data = pd.read_csv("Dataset Task 2.csv")
```

In [5]:
```
data.head()
```

Out[5]:

	id	room_id	noted_date	temp	out/in
0	__export__.temp_log_196134_bd201015	Room Admin	08-12-2018 09:30	29	In
1	__export__.temp_log_196131_7bca51bc	Room Admin	08-12-2018 09:30	29	In
2	__export__.temp_log_196127_522915e3	Room Admin	08-12-2018 09:29	41	Out
3	__export__.temp_log_196128_be0919cf	Room Admin	08-12-2018 09:29	41	Out
4	__export__.temp_log_196126_d30b72fb	Room Admin	08-12-2018 09:29	31	In

In [6]:
```
data.rename(columns={'room_id/id':'room_id'}, inplace=True)
```

In [7]:
```
data.head()
```

Out[7]:

	id	room_id	noted_date	temp	out/in
0	__export__.temp_log_196134_bd201015	Room Admin	08-12-2018 09:30	29	In
1	__export__.temp_log_196131_7bca51bc	Room Admin	08-12-2018 09:30	29	In
2	__export__.temp_log_196127_522915e3	Room Admin	08-12-2018 09:29	41	Out
3	__export__.temp_log_196128_be0919cf	Room Admin	08-12-2018 09:29	41	Out
4	__export__.temp_log_196126_d30b72fb	Room Admin	08-12-2018 09:29	31	In

```
2    Out
3    Out
4     In
```

```
In [8]: data.tail()
```

```
Out[8]:                                          id     room_id        noted_date  \
97601   __export__.temp_log_91076_7fbd08ca  Room Admin  28-07-2018 07:07
97602  __export__.temp_log_147733_62c03f31  Room Admin  28-07-2018 07:07
97603  __export__.temp_log_100386_84093a68  Room Admin  28-07-2018 07:06
97604  __export__.temp_log_123297_4d8e690b  Room Admin  28-07-2018 07:06
97605  __export__.temp_log_133741_32958703  Room Admin  28-07-2018 07:06

       temp out/in
97601    31     In
97602    31     In
97603    31     In
97604    31     In
97605    31     In
```

observation : seems like there is no use with **room_id** . so we are going to drop this column.

```
In [9]: data.drop("room_id" , inplace=True,axis = 1)
```

```
In [10]: data.tail()
```

```
Out[10]:                                          id        noted_date  temp out/in
97601   __export__.temp_log_91076_7fbd08ca  28-07-2018 07:07    31     In
97602  __export__.temp_log_147733_62c03f31  28-07-2018 07:07    31     In
97603  __export__.temp_log_100386_84093a68  28-07-2018 07:06    31     In
97604  __export__.temp_log_123297_4d8e690b  28-07-2018 07:06    31     In
97605  __export__.temp_log_133741_32958703  28-07-2018 07:06    31     In
```

```
In [11]: data.describe()
```

```
Out[11]:                temp
count  97606.000000
mean      35.053931
std        5.699825
min       21.000000
25%       30.000000
50%       35.000000
75%       40.000000
max       51.000000
```

Observation: we have maximum temperature of **51** degrees and minimum of **21** degrees.

```
In [12]: data.shape
```

```
Out[12]: (97606, 4)
```

observations: our data set has **97606** rows and **5** columns.

```
In [13]: data.columns

Out[13]: Index(['id', 'noted_date', 'temp', 'out/in'], dtype='object')

In [59]: data.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 97606 entries, 0 to 97605
Data columns (total 4 columns):
id            97606 non-null object
noted_date    97606 non-null object
temp          97606 non-null int64
out/in        97606 non-null object
dtypes: int64(1), object(3)
memory usage: 3.0+ MB
```

Observations: so totally there are 97606 rows ranging from 0 to 97605 with 5 columns. they are id , room_id, noted_date, temp, out/in.

```
In [60]: data['temp'].value_counts()

Out[60]: 39    10203
         28     8831
         29     7922
         40     7798
         31     7236
         30     6614
         37     5723
         32     5408
         27     4631
         41     4354
         36     3965
         38     3867
         42     3447
         33     3437
         34     2613
         43     2004
         44     1774
         35     1582
         45     1508
         46     1201
         47     1044
         48      971
         26      699
         49      401
         25      224
```

```
24       66
50       55
22       19
23        5
51        2
21        2
Name: temp, dtype: int64
```

Observations: we have more coulmns(10203) with **39** degrees and a couple with **21** degrees.

i am not sure, but seems like there is something wrong with this data. the column id have some encrypted values. even though you clear them, you will get id's where they were not repeated again. lets consider a scenario. i am student , and university gave me some id to enter and to exit from the room, in this case id's must be repeated in the column with enter and exit times. but this didn't happen in your dataset. please correct me if iam wrong.