Centro Universitário Instituto de Educação Superior de Brasília
Departamento de Pós graduação
Especialização em Big Data, BI e Analytics Aplicados aos Negócios

Projeto Integrador I

Aluna: Mariana Borges de Sampaio
Professor: William de Almeida Silva
Disciplina: Fundamentos de Big Data e Conhecimentos de Dados

Agosto
2023

Centro Universitário Instituto de Educação Superior de Brasília
Departamento de Pós graduação
Especialização em Big Data, BI e Analytics Aplicados aos Negócios

Projeto Integrador I

Primeiro relatório do Projeto Integrador 1 da disciplina de Fundamentos de Big Data e Conhecimentos de dados.

Aluna: Mariana Borges de Sampaio

Professor: William de Almeida Silva

Agosto
2023

Conteúdo

1 Introdução

Ao longo dos anos as empresas de todos os países e de todos os continentes foram se desenvolvendo, cada um em sua área específica, sejam empresas de corretoras de seguro, empreiteiras, empresas de tecnologia, empresas de saúde, entre os mais diversos tipos e variados setores em que uma empresa pode se aplicar. Apesar dessas empresas terem conceitos e fundamentos diferentes visto que elas são pertencentes a mundos diferentes, tem-se detém um conjunto de dados, esse dado muitas vezes pode não ser estruturado, sendo ele um dado bruto, um dado que só teria sentido se fosse tratado a fim de se tornar uma informação.

Sendo assim, cada empresa existente até os dias de hoje inevitavelmente possui dado, logo possui informação. Esse dado pode ser ele armazenado em um banco de dados estruturado ou pode ser um big data, sendo este um dado não estruturado. O big data tem por definição ser um conjunto de dados que possuem uma maior variedade, maior volume e estes dados têm uma maior velocidade. Sendo assim, o seu armazenamento é diferente do que um dado estruturado. Para isso conforme foram-se passando os anos foi sendo necessário nos meios do trabalho ter profissionais que são voltados para essa área.

2 Dicionário de dados e Análise de dados

2.1 Dicionário de dados

A fonte de dados em questão foi extraída do Kaggle, conforme o que foi passado na dsecrição da atividade ativa.

Ao analisar o documento pode-se perceber a existência das seguintes colunas:

- id;
- roomid/id;
- noteddate;
- temp;
- out/in.

A seguir, tem-se a coluna com seu respectivo significado.

- id - corresponde à um valor alfanumérico que indica uma identificação única, sendo assim cada linha é um momento em que foi registrada a temperatura do edifício;
- roomid/id - corresponde ao nome da sala do edifício em que foi medida a temperatura. Para a identificação das medidas, tem-se o id que é um alfanumérico, visto que esses casos foram todos registrados em momentos aleatórios.
- noteddate - corresponde a data de anotação que foi identificada a temperatura no edifício, sendo está registrada em intervalos aleatórios;
- temp - corresponde a temperatura que foi atingida no edifício;
- out/in - indica se a classificação de cada se a temperatura medida foi no exterior ou no interior do edifício.

2.2 Análise de dados

Para realizar a análise de dados, foi necessário entender um pouco mais sobre o contexto do arquivo que foi extraído do kaggle. Os dados apresentados correspondem a dados oriundos de um sistema de monitoramento de temperatura da sala do edifício empresarial (admin), tanto no exterior como no interior do edifício.

Para realizar a análise dos dados foi feita a leitura do que foi apresentado a fim de entender o contexto da análise. A partir disso, foi desenvolvido um código no arquivo (ipynb) jupyter notebook que possui todo o código desenvolvido que foi criado com sessão do aplicativo juptyer notebook. Para isso, utilizei o vscode studio com a extensão referente ao juptyer notebook. Para manter o versionamento, utilizei o github, dessa forma, conforme avançava na análise e no código realizada um update no github que foi destinado a esse trabalho,https://github.com/sampaiomariana/iot-devices-data-analysis.Para o desenvolvimento do arquivo de relatório, utilizei o overleaf, que também conforme avançava foi sendo atualizado no github.

Na análise dos dados, foram seguidos cinco passos durante o desenvolvimento, sendo estes, os seguintes:

- Importar os dados;
- Visualizar a base;
- Tratamento de erros;
- Análise inicial dos dados;
- Análise profunda dos dados.

Para realizar todos esses passos, foram utilizadas as bibliotecas pandas e statistics. Para realizar a primeira análise e entender o os dados, importei os dados utilizando a biblioteca pandas, dessa forma consegui ver as colunas e o conteúdo das colunas, a partir dessa primeira análise, foi possível ver que existem 97606 linhas e 5 colunas, sendo assim, tem-se 97606 registros de temperatura do edifício em momentos aleatórios.

```
import pandas as pd

#Lendo a base de dados do IOT-temp do kaggle;
database = pd.read_csv("IOT-temp.csv")
display(database)
```

[17]

...

	id	room_id/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
...	...	...	...	...	...
97601	__export__.temp_log_91076_7fbd08ca	Room Admin	28-07-2018 07:07	31	In
97602	__export__.temp_log_147733_62c03f31	Room Admin	28-07-2018 07:07	31	In
97603	__export__.temp_log_100386_84093a68	Room Admin	28-07-2018 07:06	31	In
97604	__export__.temp_log_123297_4d8e690b	Room Admin	28-07-2018 07:06	31	In
97605	__export__.temp_log_133741_32958703	Room Admin	28-07-2018 07:06	31	In

97606 rows × 5 columns

A fim de entender qual era o tipo de conteúdo em cada linha, foi feito um comando com uma descrição do conteúdo de cada linha. Essa análise é feita para ver se existem campos nulos nas colunas, caso existam eles devem ser tratados. Nesse caso, os tipos de conteúdo eram int64 e object e não existem campos nulos nas colunas. Visto que existiam duas colunas que representam a chave primária da linha, sendo ela, o id e a roomid/id, optei por retirar a coluna id.

```python
#Para visualizar os tipos de dados;
#Vendo a base o que podemos observar que todas as linhas estão preenchidas, não sendo necessário retirar dados nulos;
#Apenas a coluna int64 é inteiro, enquanto os demais são object;
display(database.info())

#Como a coluna id indica apenas a quantidade de ids registrados e essa contagem já será feita, essa informação pode ser retirada.
# linha -> axis = 0
# coluna -> axis = 1
database = database.drop("id", axis = 1)
display(database.info())
```

[18]

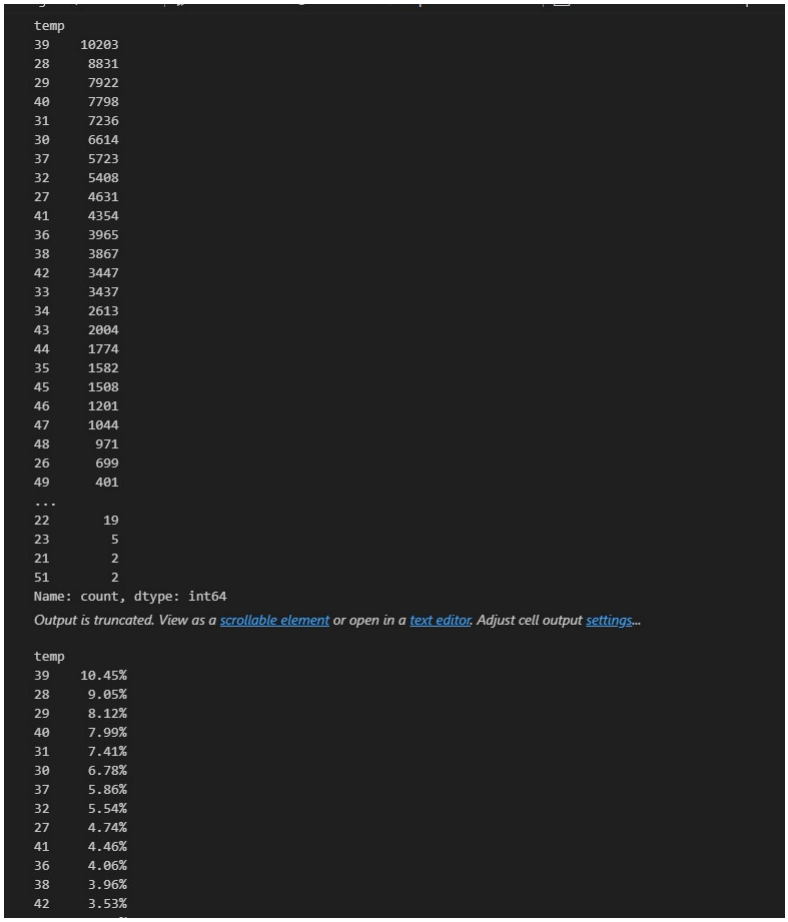
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 97606 entries, 0 to 97605
Data columns (total 5 columns):
 #   Column      Non-Null Count  Dtype 
---  ------      --------------  ----- 
 0   id          97606 non-null  object
 1   room_id/id  97606 non-null  object
 2   noted_date  97606 non-null  object
 3   temp        97606 non-null  int64 
 4   out/in      97606 non-null  object
dtypes: int64(1), object(4)
memory usage: 3.7+ MB

None

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

None
```

Após isso, realizei uma análise focada em cada coluna a fim de verificar na coluna do roomid/id se existia algum campo diferente de roomid, no caso da coluna de noteddate, a fim de verificar a quantidade de registros em uma data e o seu percentual. Foi possível observar que na data de 12/09/20218 tiveram mais medições de temperatura, sendo ela 65 registros, o que corresponde a 0.07 percentual do total de registros.

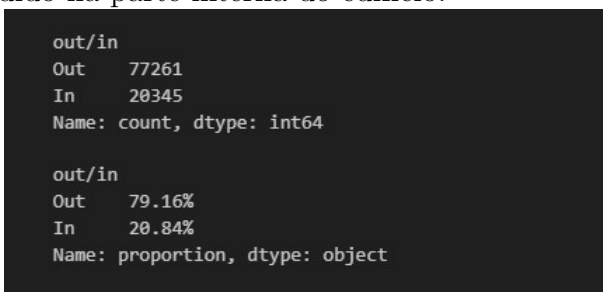
```python
#Verificar a quantidade dos tipos de room/id para verificar se existe algum grupo além de room admin
display(database["room_id/id"].value_counts())
#Verificar os tipos de noted_date para verificar e respectivamente a sua quantidade e o seu percentual
display(database["noted_date"].value_counts())
display(database["noted_date"].value_counts(normalize=True).map("{:.2%}".format))
#Verificar os tipos de temp para verificar e respectivamente a sua quantidade e o seu percentual
display(database["temp"].value_counts())
display(database["temp"].value_counts(normalize=True).map("{:.2%}".format))
#Verificar os tipos de out/in para verificar e respectivamente a sua quantidade e o seu percentual
display(database["out/in"].value_counts())
display(database["out/in"].value_counts(normalize=True).map("{:.2%}".format))
```

[24] Python

O mesmo foi feito com temp, que corresponde a temperatura, nessa análise foi possível observar que houve 10203 registros com temperatura por 39 graus, que corresponde a 10.45 percentual.



Também para a coluna de out/in foi feito, e foi possível observar que a maioria dos registros foram feitos na parte externa do edifício, pela análise, tem-se que 79.16 percentual foi medido externamente e 20.84 percentual foi medido na parte interna do edifício.



Para ter uma análise mais precisa, realizei o agrupamento dos dados, sendo baseado na coluna out/in e extrai a média da temperatura registrada, dessa forma observei que a temperatura média no lado externo do edifício foi de 36.2 graus e que no lado interno do edifício foi de 30.45 graus.

```
import statistics
from statistics import mean

#Média de temp de acordo com out/in
display(database.groupby("out/in").mean(numeric_only=True))
#Média de temp de acordo com noted_date
display(database.groupby("noted_date").mean(numeric_only=True))
```

out/in	temp
In	30.453920
Out	36.265244

noted_date	temp
01-11-2018 00:06	32.0
01-11-2018 00:07	40.0
01-11-2018 00:09	39.0
01-11-2018 00:13	40.0
01-11-2018 00:23	40.0
...	...
31-10-2018 23:52	40.0
31-10-2018 23:53	39.0
31-10-2018 23:54	32.0
31-10-2018 23:55	40.0
31-10-2018 23:57	39.0

27920 rows × 1 columns

Fazendo esse mesmo agrupamento de dados, porém utilizando a coluna de noteddate como base, tem-se que a temperatura média em cada data, sendo assim, observa-se que a temperatura variou entre 32 e 40 graus durante esses dias.

3 Considerações finais

Bibliografia

AGUIRRE, L. A. Introdução à Identificação de Sistemas, Técnicas Lineares e Não lineares Aplicadas a Sistemas Reais. Belo Horizonte, Brasil, EDUFMG. 2004.

Anexo