Capstone Project

Daniela Ribeiro
Nathalia Guerra

Data exploration and visualization:

- Significant time was dedicated to this part of the study.

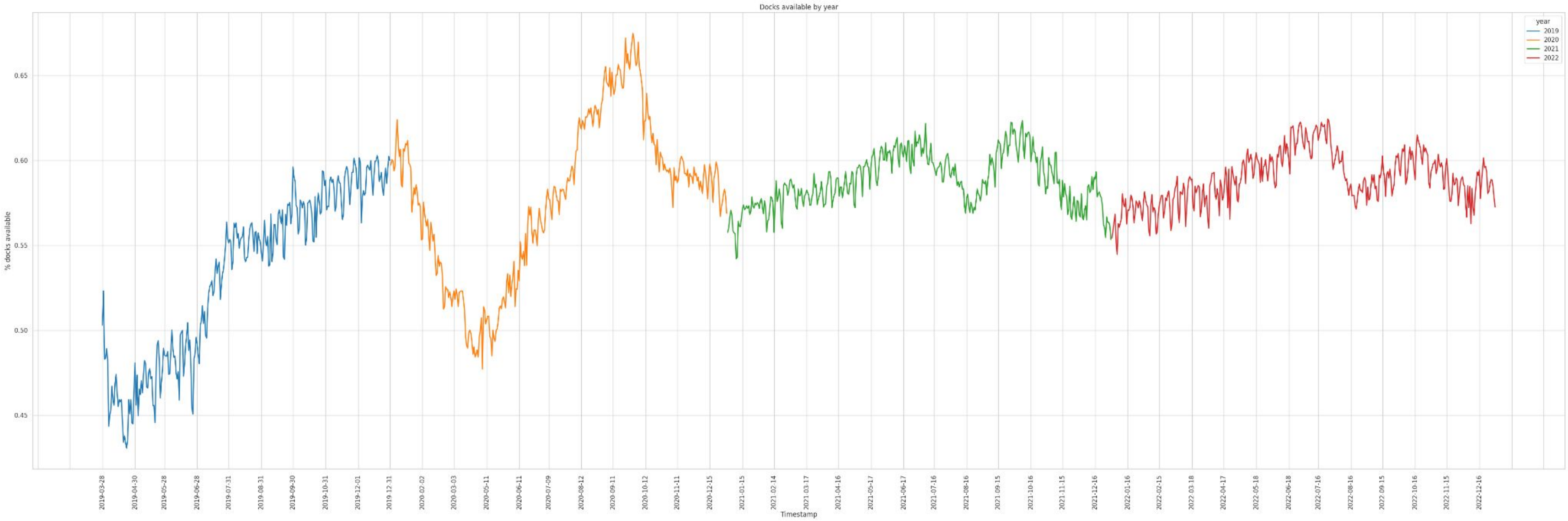
- Understanding, fixing bugs, load and process the data.

```
[ ] df_station_info
```

	station_id	name	physical_configuration	lat	lon	altitude	address	post_code	capacity	is_charging_station	nearby_distance	_ride_code_support	rental_uris	cro
0	1	GRAN VIA CORTS CATALANES, 760	ELECTRICBIKESTATION	41.397978	2.180107	16.0	GRAN VIA CORTS CATALANES, 760	08013	46	True	1000.0	True	None	
1	2	C/ ROGER DE FLOR, 126	ELECTRICBIKESTATION	41.395488	2.177198	17.0	C/ ROGER DE FLOR, 126	08013	29	True	1000.0	True	None	
2	3	C/ NÀPOLS, 82	ELECTRICBIKESTATION	41.394156	2.181331	11.0	C/ NÀPOLS, 82	08013	27	True	1000.0	True	None	
3	4	C/ RIBES, 13	ELECTRICBIKESTATION	41.393317	2.181248	8.0	C/ RIBES, 13	08013	21	True	1000.0	True	None	
4	5	PG. LLUIS COMPANYS, 11 (ARC TRIOMF)	ELECTRICBIKESTATION	41.391103	2.180176	7.0	PG. LLUIS COMPANYS, 11 (ARC TRIOMF)	08018	39	True	1000.0	True	None	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	
504	515	C/ SANT ADRIÀ, 43	ELECTRICBIKESTATION	41.435207	2.194800	19.0	C/ SANT ADRIÀ, 43	08030	24	True	1000.0	True	None	
505	516	C/ SANT ADRIÀ, 88	ELECTRICBIKESTATION	41.435460	2.200157	15.0	C/ SANT ADRIÀ, 88	08030	21	True	1000.0	True	None	
506	517	AV. RASOS DE PEGUERA, 10	ELECTRICBIKESTATION	41.462095	2.178959	44.0	AV. RASOS DE PEGUERA, 10	08033	20	True	1000.0	True	None	

Data exploration and visualization:

Docks available (%)

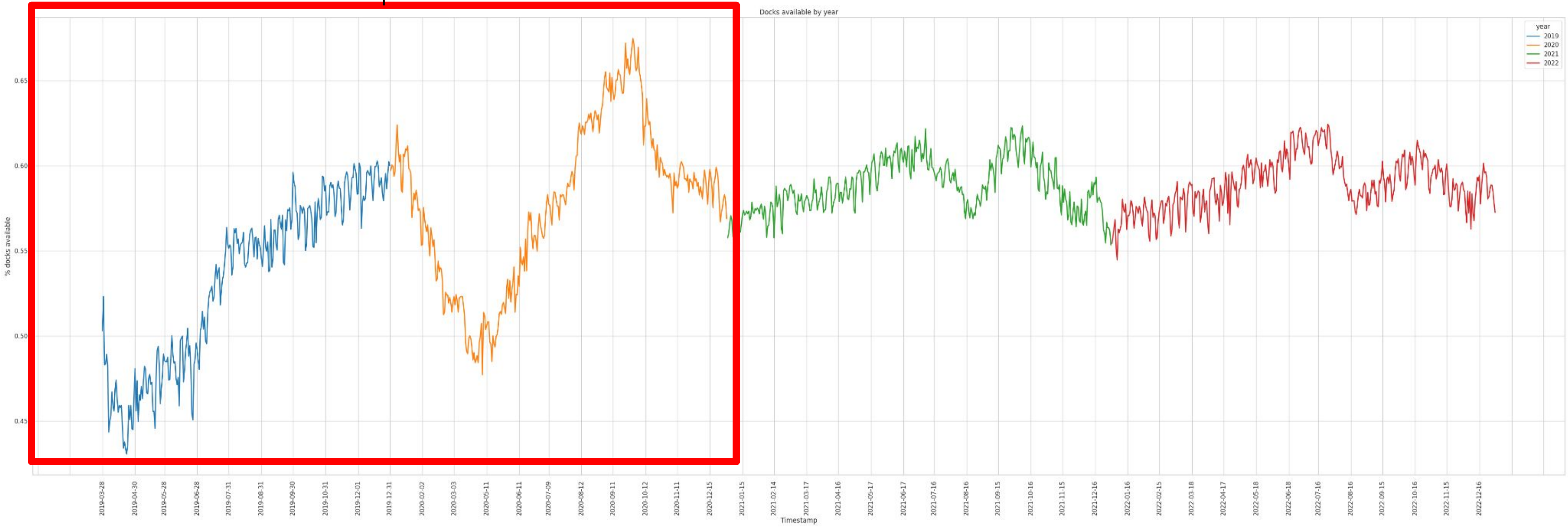


Timestamp

Data exploration and visualization:

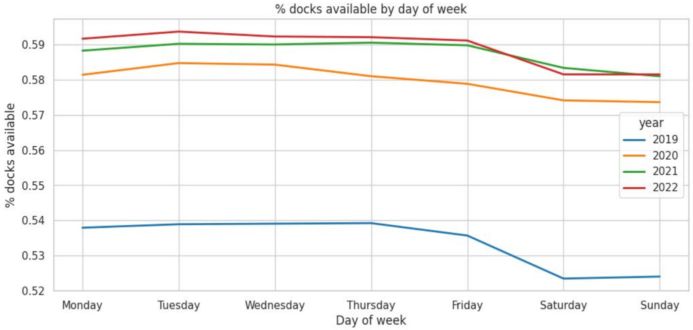
First year of service and Covid behaviour

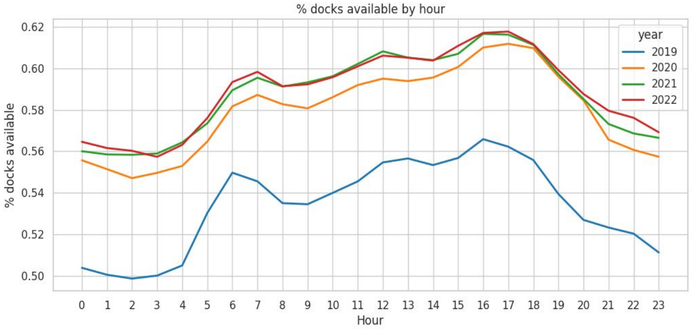
Docks available (%)

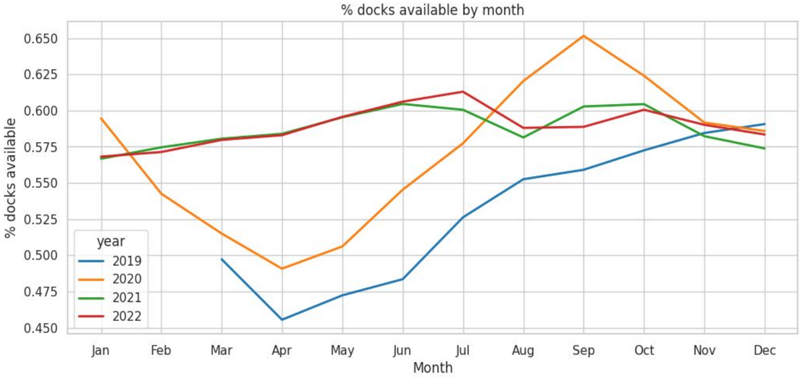


Timestamp

Data exploration and visualization:

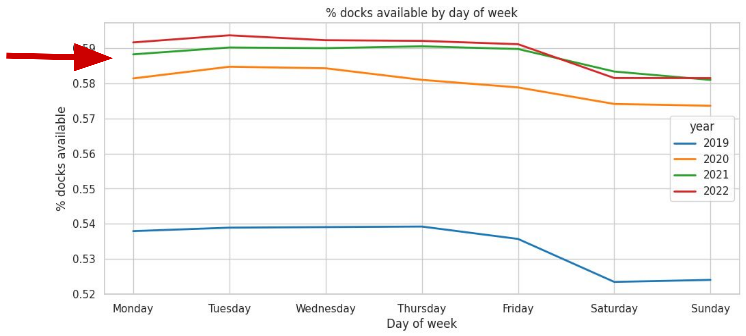


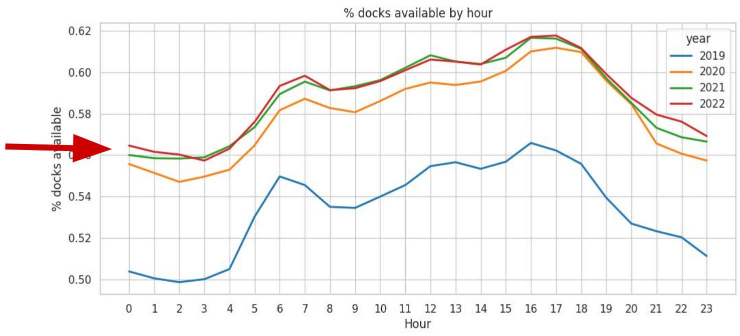


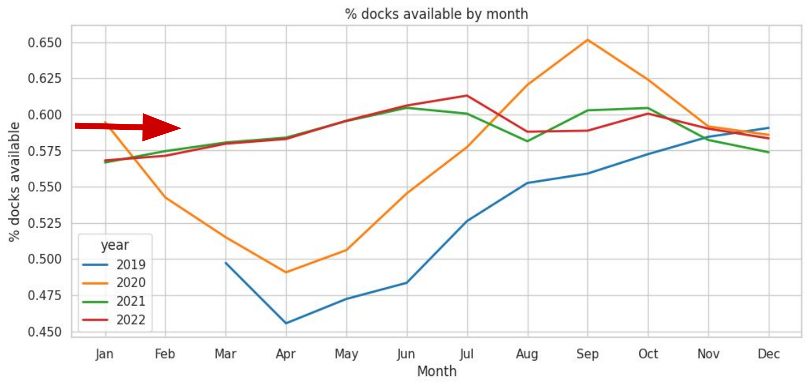


Data exploration and visualization:

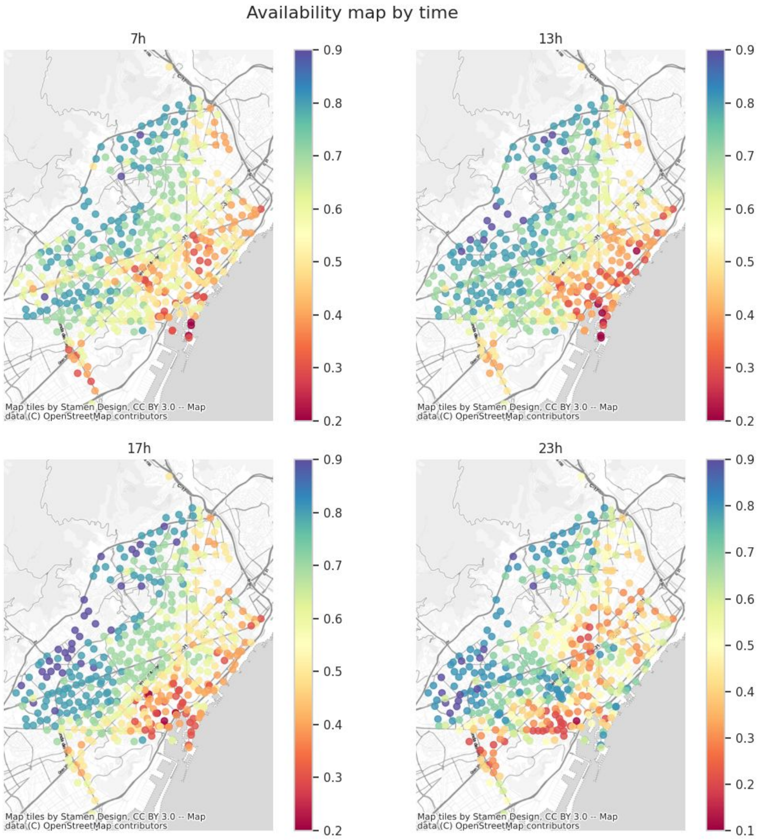
The team decided to consider only 2021-2022 data for analysis.







Data exploration and visualization:



Availability throughout the city: 7h, 13h, 17h and 23h.

Correlation: Availability ↑

Altitude ↑

Correlation and variables:

```
  corr_matrix2 = df_model.corr()
percentage_docks_available    1.000000
ctx-1                         0.910030
ctx-2                         0.791951
ctx-3                         0.689114
ctx-4                         0.597569
altitude                      0.342513
lat                           0.140444
station_id                    0.138987
label                         0.047892
temp                          0.036678
hour                          0.031924
month                         0.018863
post_code                     0.016895
year                          0.002786
day                           0.001416
wind                         -0.000413
prec                         -0.008043
capacity                     -0.009374
day_of_week                  -0.012123
lon                          -0.237852
Name: percentage_docks_available, dtype: float64
```

High correlation does not mean high accuracy for the model.

Diverse sets of categorical and numerical variables were tested in both models:

- Regression Models
- Deep Learning network

Regression model:

The following models were explored, evaluated, improved, and analyzed based on data correlation and results.

- LinearRegression()
- DecisionTreeRegressor()
- RandomForestRegressor()
- MLPRegressor()
- GradientBoostingRegressor()
- KNeighborsRegressor()

```
Model: LinearRegression()
CV score: 0.11001660346922633
Model: DecisionTreeRegressor()
CV score: 0.1533719138121837
Model: RandomForestRegressor()
CV score: 0.10599045759081109
Model: MLPRegressor()
CV score: 0.10640971043576815
Model: GradientBoostingRegressor()
CV score: 0.10719596402954475
Model: KNeighborsRegressor()
CV score: 0.13866112971310357
```

RandomForestRegressor presented best results.

Numerical variables considered: ['label', 'prec'] (label from clustering, and precipitation data)

Categorical variables (OneHotEncoding): ['altitude', 'lat', 'lon', 'hour', 'month', 'day_of_week', 'station_id'] and 'contexts'.

Deep learning model:

Categorical variables do not present significant effect on learning process.

The group decided to try Deep Learning model to verify embedding impact on categorical variables.

The Keras functional API was used to build the model -> more flexible than the keras.Sequential API (multiple inputs)

- Different variables and model structures were tested.

- Fair results were obtained considering the following

categorical variables: ['station_id', 'hour']

Numerical variables: ['ctx-4', 'ctx-3', 'ctx-2', 'ctx-1']

Deep learning model:

```
[48] # Input data dimensions
     input_data = [cat_inputs, numeric_inputs]
     input_data
```

```
[<KerasTensor: shape=(None, 4) dtype=float32 (created by layer 'cat_inputs')>,
 <KerasTensor: shape=(None, 5) dtype=float32 (created by layer 'numeric_inputs')>]
```

```
[49] #Concatenate embedding layers
     emb_data = tf.keras.layers.Concatenate(axis=-1, name="concat_layer")([emb_cat, emb_numeric])
     emb_data
```

```
<KerasTensor: shape=(None, 337) dtype=float32 (created by layer 'concat_layer')>
```

```
#Non-sequential model
tf.keras.backend.clear_session()
x = tf.keras.layers.Dense(16, activation='relu')(emb_data)
x = tf.keras.layers.Dense(10, activation='relu')(x)
x = tf.keras.layers.Dense(2, activation='relu')(x)
x = tf.keras.layers.Dense(1)(x)
model = Model(inputs=input_data, outputs=x)
model.summary()
```

Gràcies!

Obrigado!

https://github.com/NathaliaBatalha/CapstoneProject_NathaliaDaniela