Seoul Bike Sharing Demand

Inès PEREZ & Mathilde SALAÜN

DIA 5

Overview

- **Exploring Data**: Importation, NaN values, definition of new variables
- **Visualisation**: Ploting : matplotlib, seaborn, plotly express, bokeh
- **Pre-processing**: Encoding
- **Modeling**: Regression and Classification models, comparaison
- **Conclusion**: Best prediction and best model, what can we say ?
- **API**: Flask

Exploring our dataset

WHAT WE HAD

- Date : year-month-day
- **Rented Bike count - Count of bikes rented at each hour**
- Hour - Hour of the day
- Temperature-Temperature in Celsius
- Humidity - %
- Windspeed - m/s
- Visibility - 10m
- Dew point temperature - Celsius
- Solar radiation - MJ/m2
- Rainfall - mm
- Snowfall - cm
- Seasons - Winter, Spring, Summer, Autumn
- Holiday - Holiday/No holiday
- Functional Day - NoFunc(Non Functional Hours), Fun(Functional hours)

OUR GOAL

- Study the impact of all the variables on the number of rented bikes

Exploring our dataset

2 TYPES OF VARIABLES

- **Temporal ones :**

Date

Hour

Holiday

Functional Day

- **Meteorological ones :**

Temperature

Humidity

Windspeed, Visibility

Dew point temperature

Solar radiation

Rainfall

Snowfall

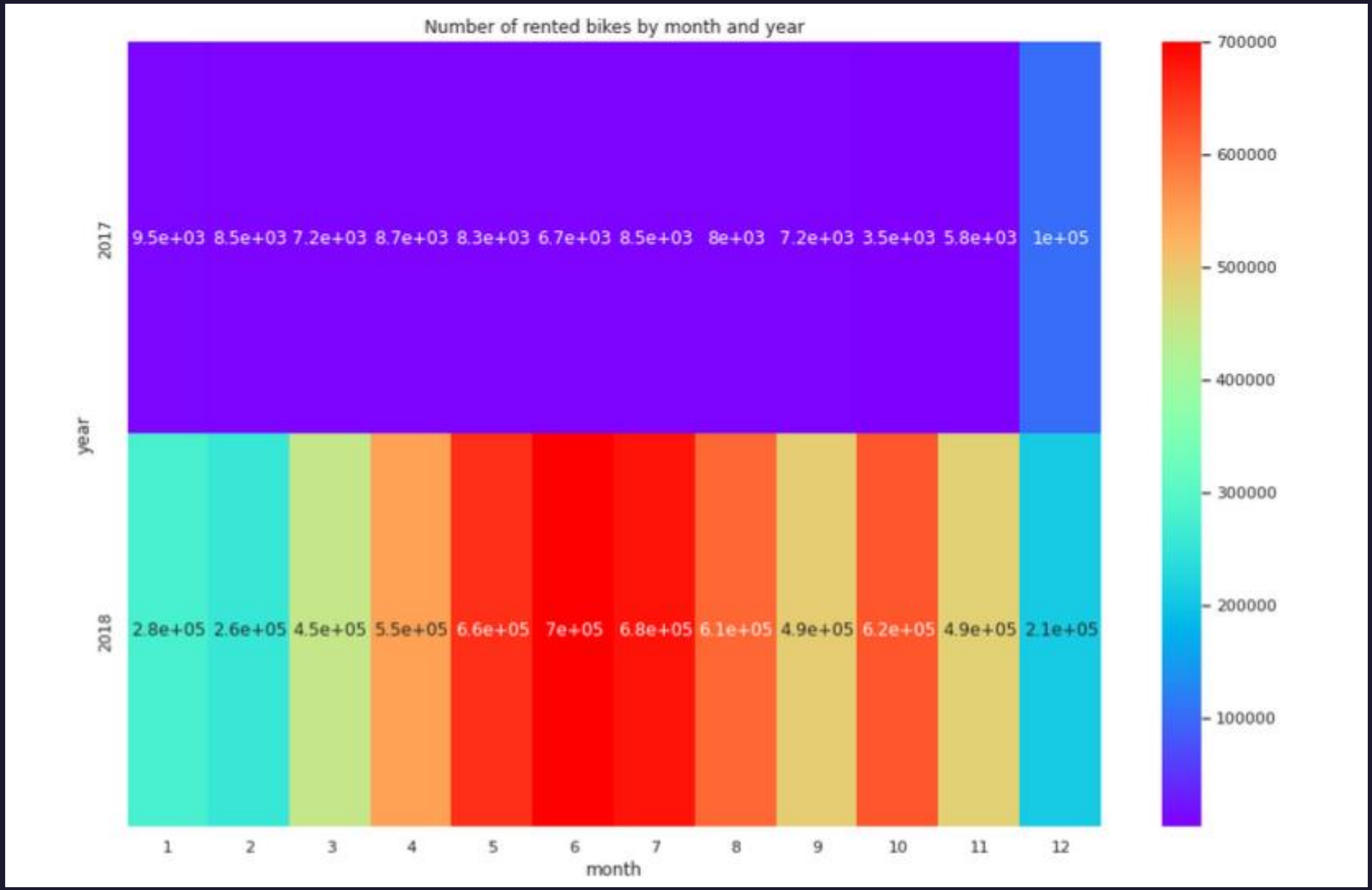
Seasons

WHAT WE HAVE DONE

- Fixing our target : Rented Bike count
- No NaNs values
- Rename columns
- New variables : day, month, year, Moment_of_day and bike_affluence
- Correction of the season label

Visualisation

TEMPORAL VARIABLES

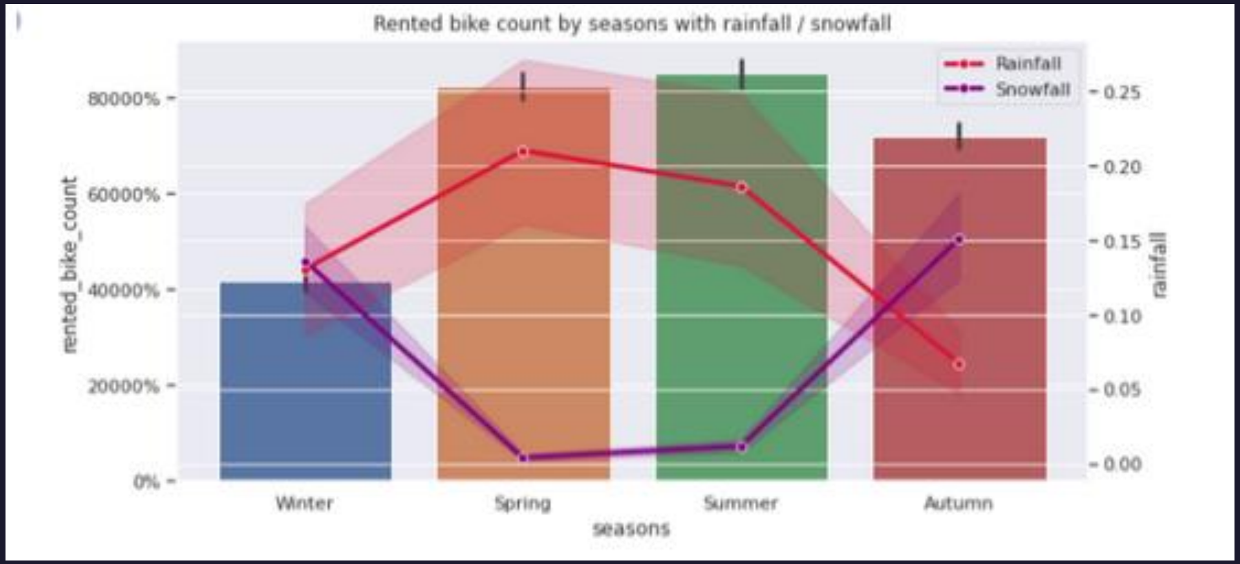


Visualisation

METEOROLOGICAL VARIABLES

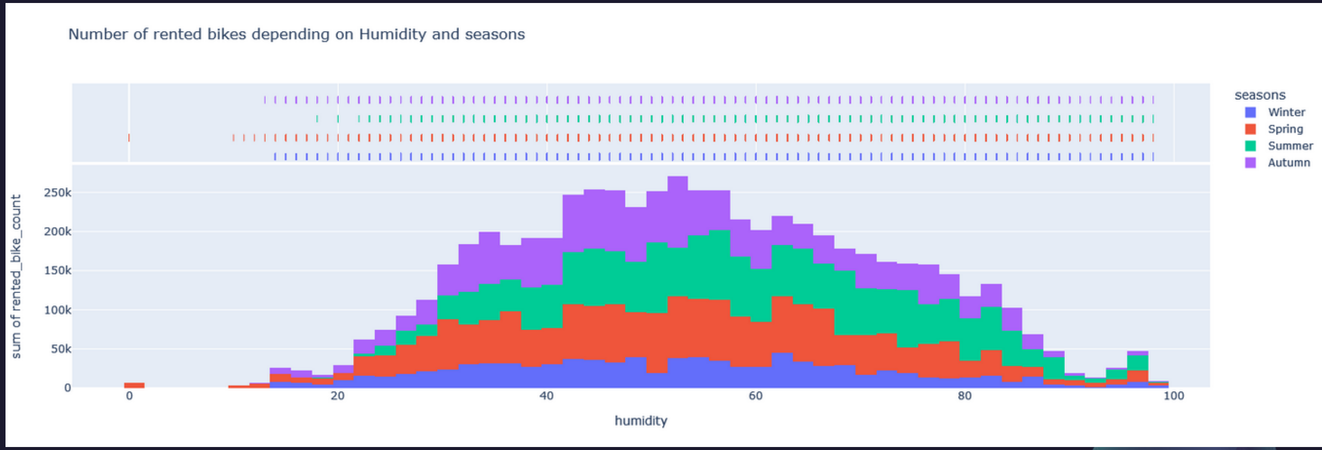
- No matter much : Temperature, Humidity, Dew point temperature, Rainfall, Snowfall

- Matter : Wind speed, Visibility, Solar radiation



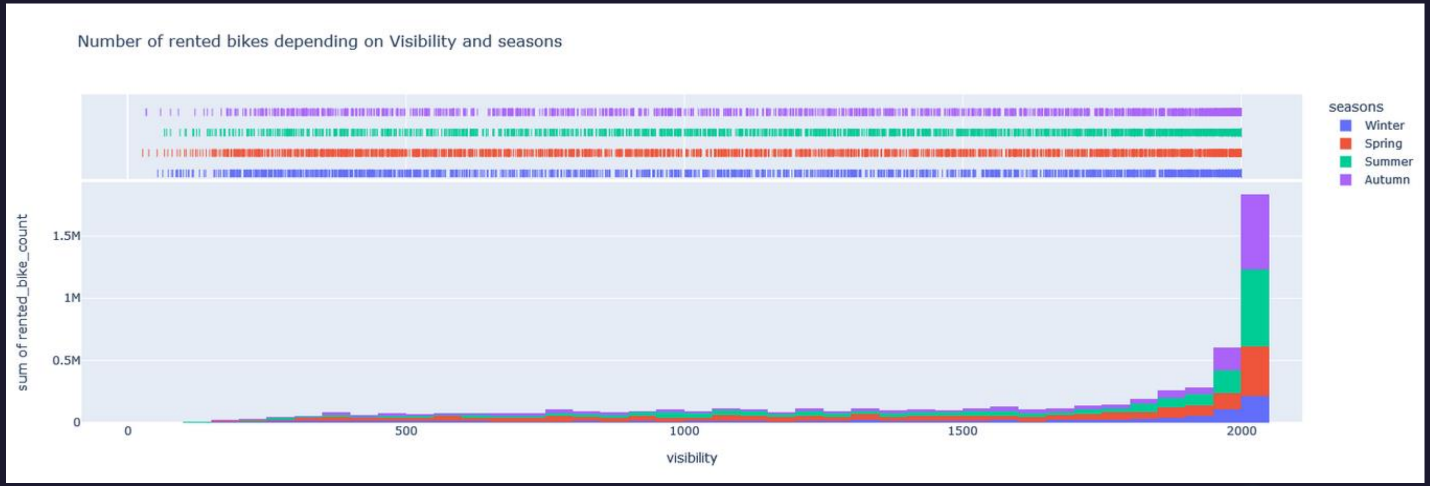
Visualisation

METEOROLOGICAL VARIABLES



- No matter : Temperature, Humidity, Dew point temperature, Rainfall, Snowfall

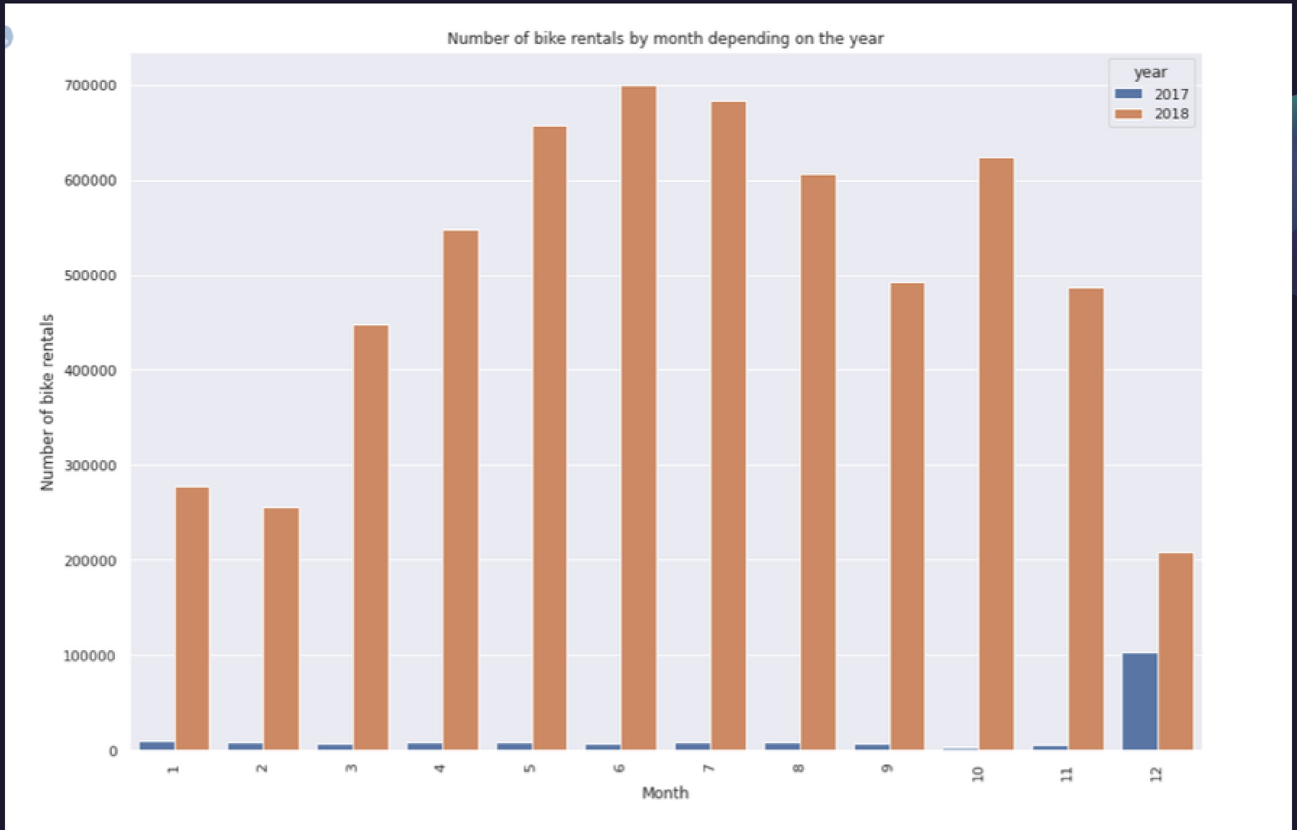
- Matter : Wind speed, Visibility, Solar radiation



Pre-processing

DROP

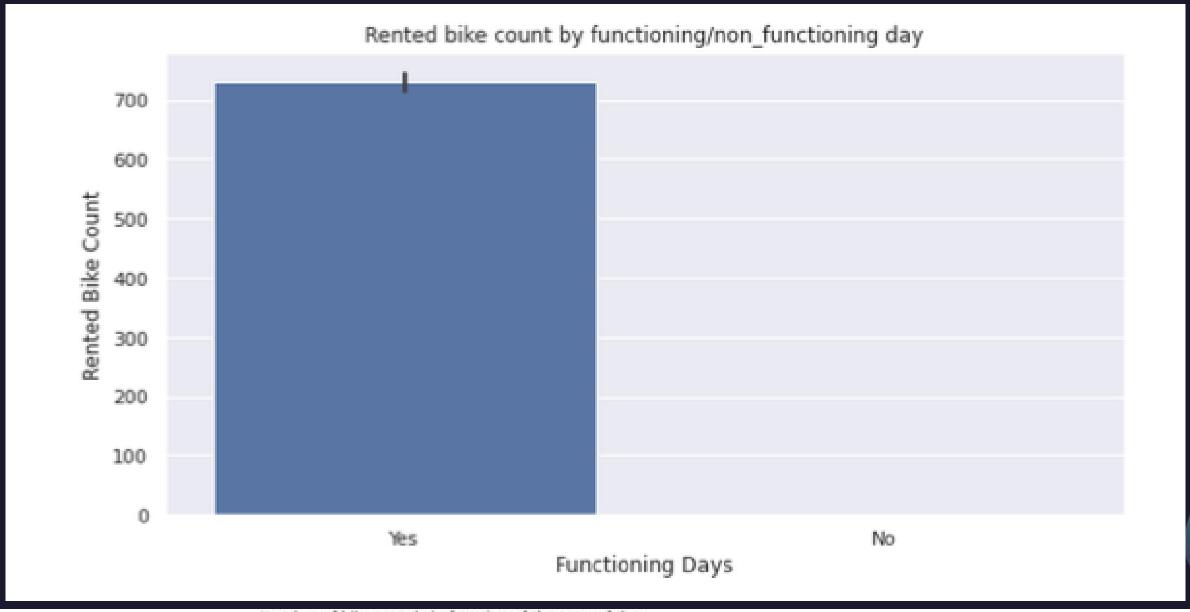
- Moment_of_day
- Functioning_day
- 2017 year

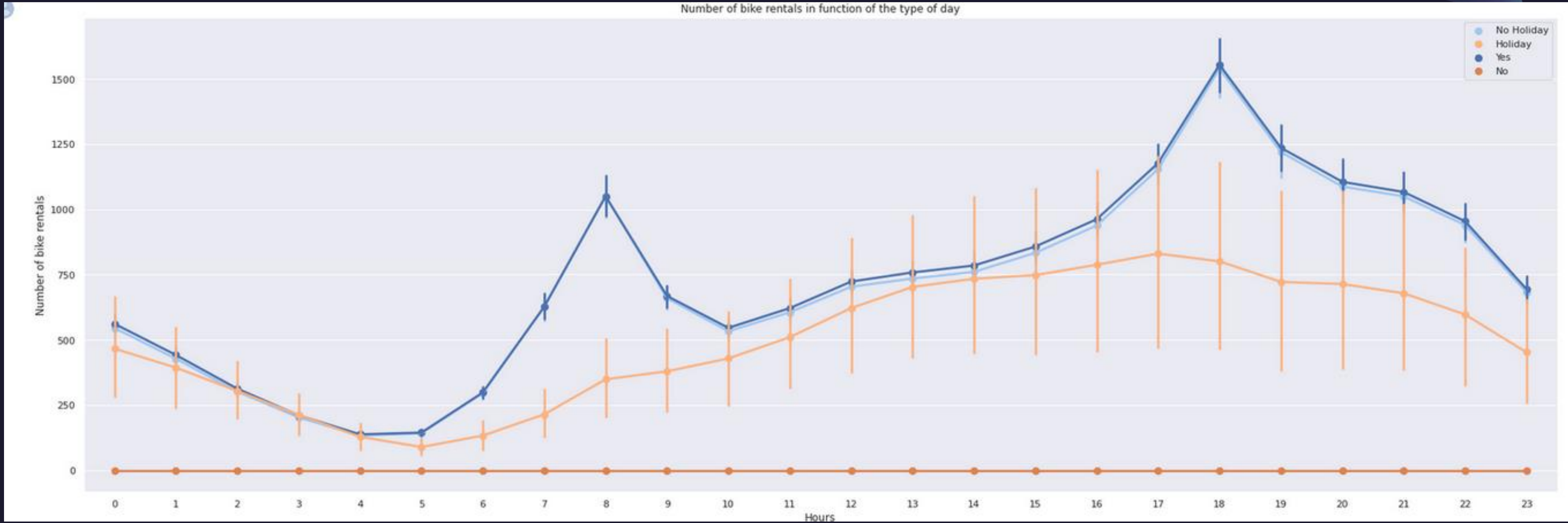


Pre-processing

DROP

- Moment_of_day
- Functioning_day
- 2017 year



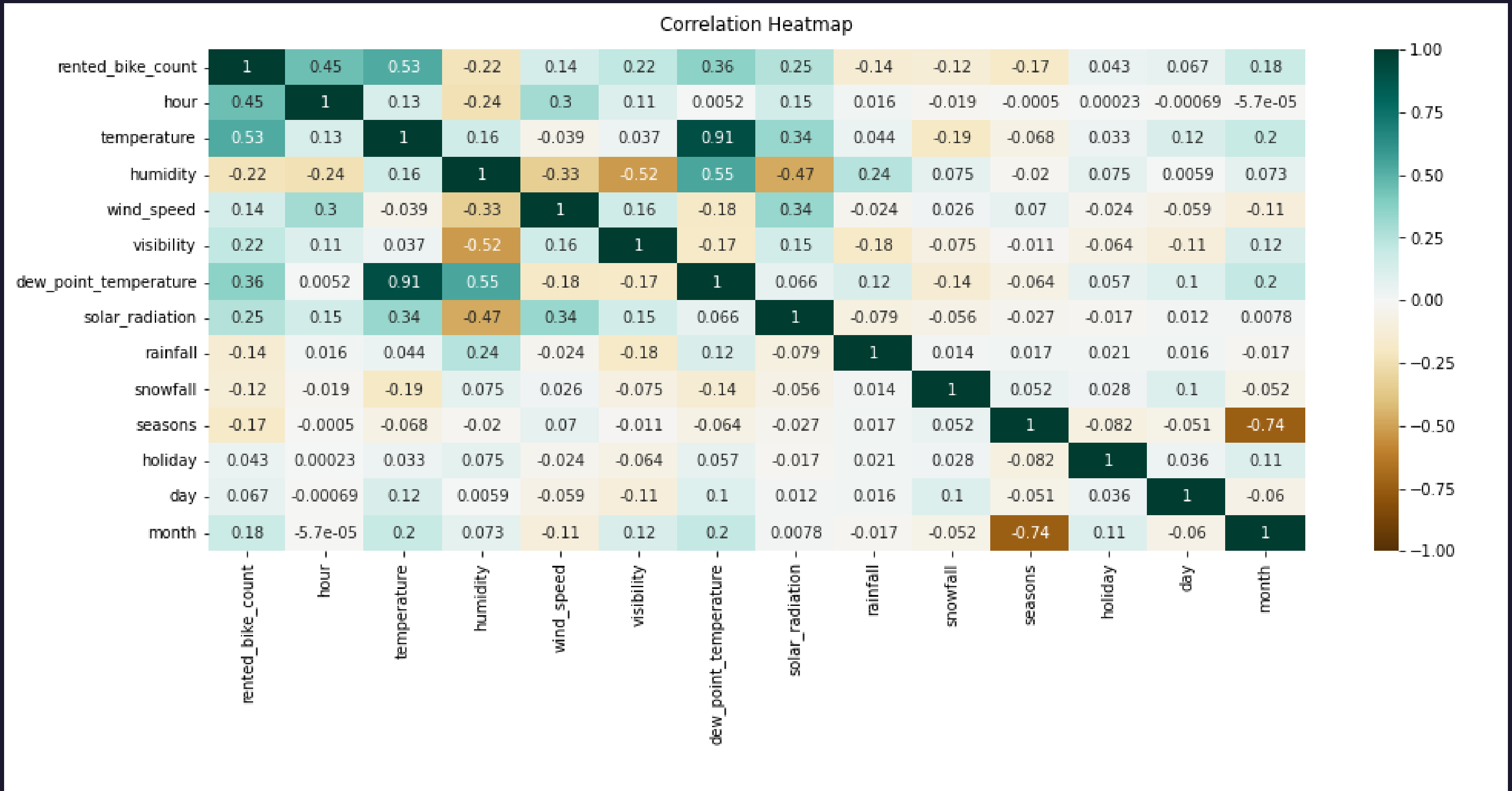


Pre-processing

CORRELATION BETWEEN VARIABLES

- Temperature : 0.53
- Hour : 0.45
- Dew point temperature : 0.36

-> but extremely correlated with temperature (0.91)



Pre-processing

ENCODER

- Seasons

BINARIZER

- Holiday

NORMALIZE

- All predicators for classification

Modeling

REGRESSION MODELS

- KNN Regressor
- Hist Gradient Boosting Regressor
- Gradient Boosting Regressor
- Bagging Regressor
- Extra Trees Regressor
- Random Forest Function
- LGBM Regressor

CLASSIFICATION MODELS

- LGBM Classifier
- Random Forest
- Extra Trees
- KNN Classifier

Modeling : Comparison - Accuracy

	Regressor	Classifier
Multiple Linear Regression	50.78%	-
KNN	57.52%	58%
Gradient Boosting	86.59%	-
Bagging	86.76%	-
Random Forest	87.38%	78.25%
Extra Trees	88.25%	78.35%
LGBM	88.60%	78.82%
Hist Gradient Boosting	**89.77%**	-

Conclusion

- Our best model is :

Hist Gradient Boosting Regressor

(HGBR) ~89.77%

- Our second best model is :

Light Gradient Boosting Machine Regressor

(LGBMR) ~88.60%

- Hour and Temperature are the two variables that influe the most on the number of rented bikes

API - Flask

LGBMR model