TEAM D – PYTHON GROUP PROJECT

BIKE SHARE USERS IN WASHINGTHON D.C.

Umut Varol, Dennis Pedersen, Milan Medina, Tony Matta,
Vimal Ramakrishnan, Alexandre Bouamama, Eugen Wettstein

DATA INFORMATION AND AIM

Dataset from Kaggle

Washington D.C.

2011 - 2012

Prediction of the hourly bikeshare rental

EDA

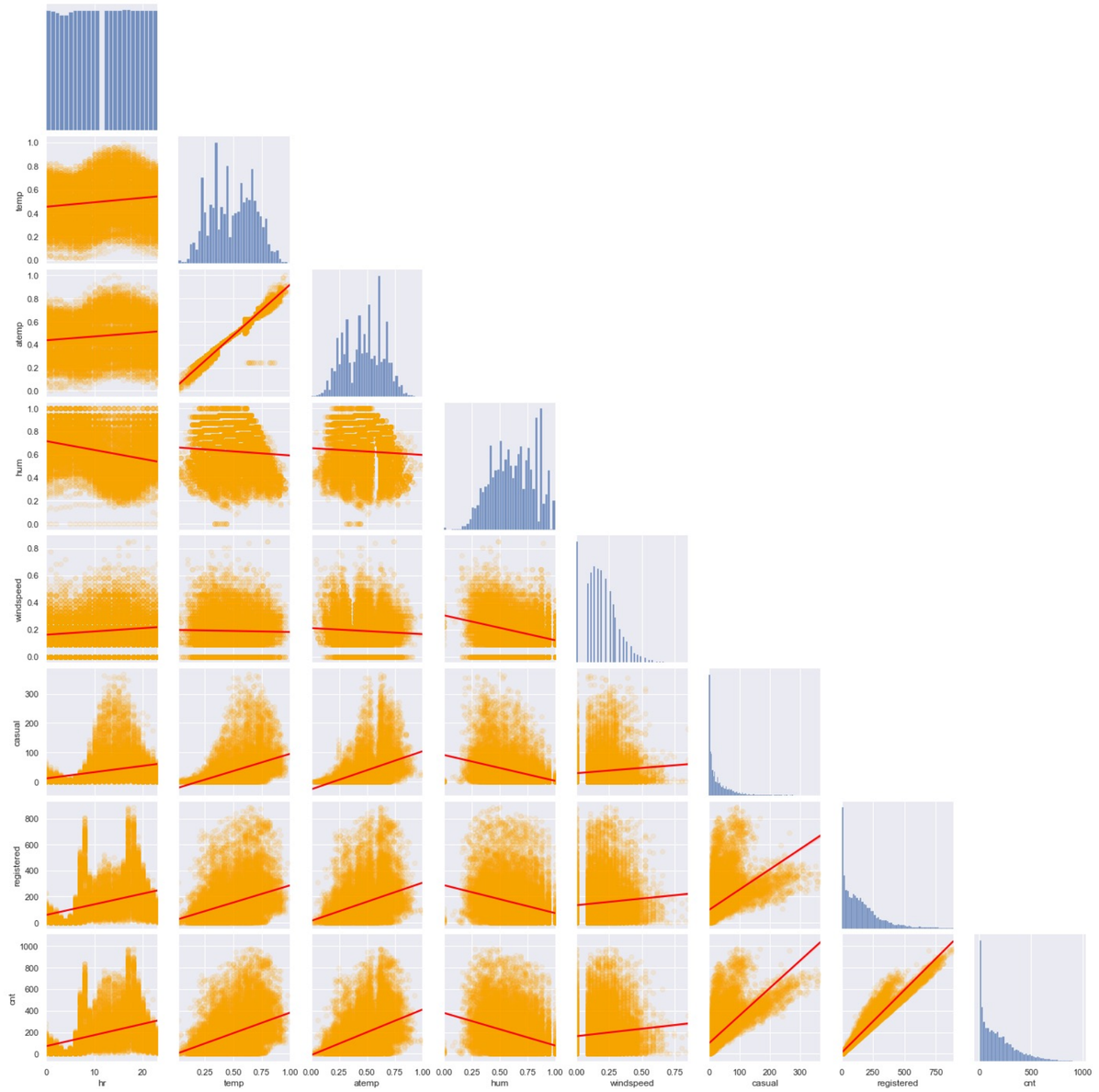
FACTS ABOUT THE DATA

VARIABLES IN THE DATASET: 11

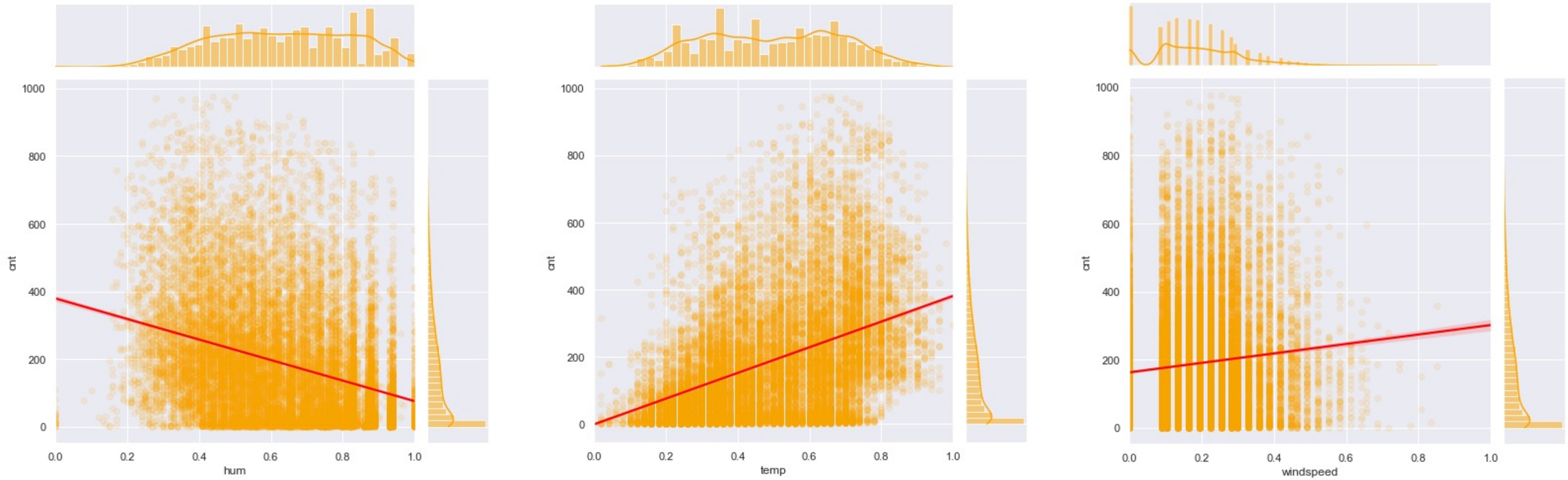
TOTAL NUMBER OF OBSERVATIONS: 17379

MISSING VALUES: 499

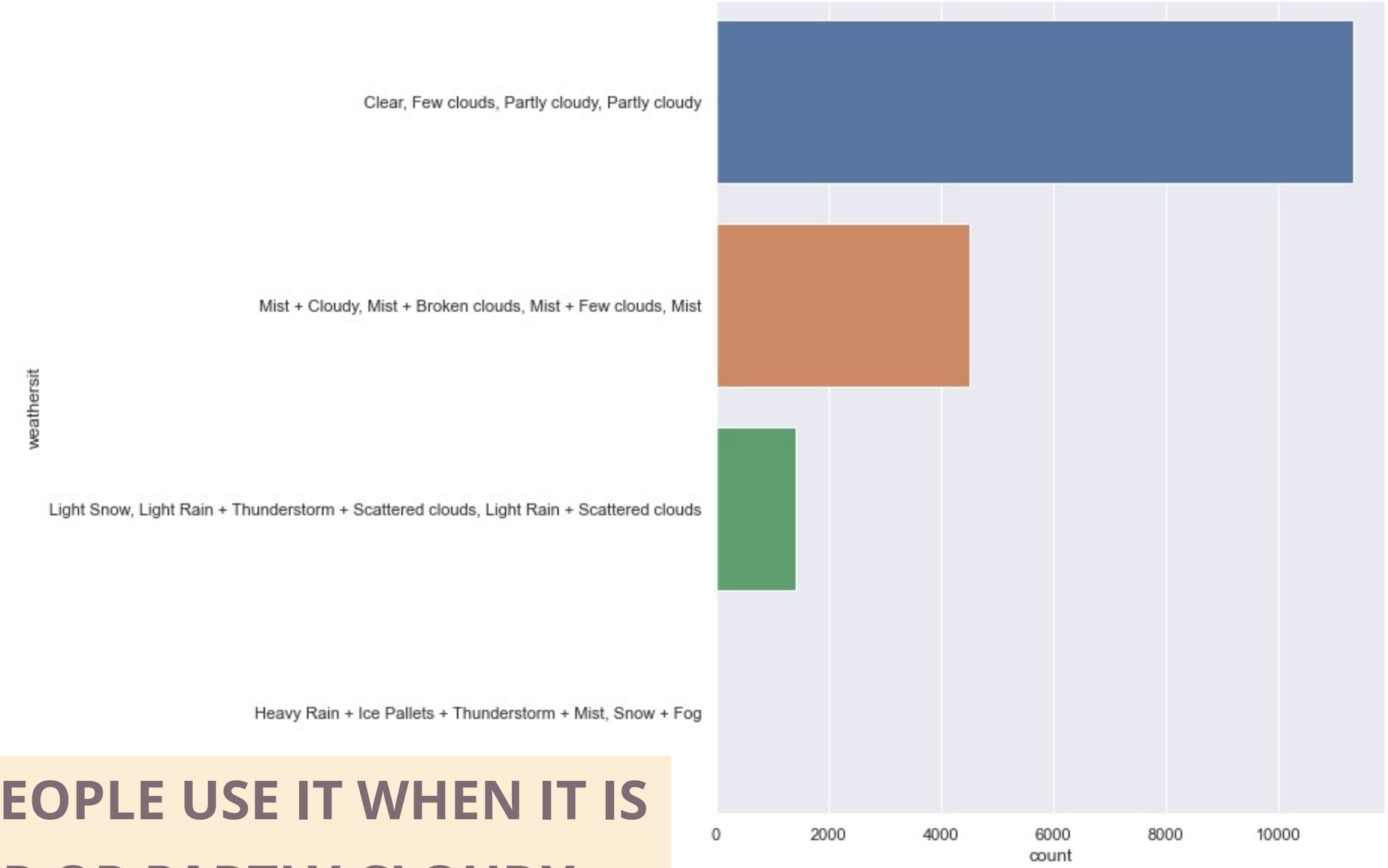
DATATYPES: 9 NUMERICAL, 1 DATETIME, 1 CATEGORICAL



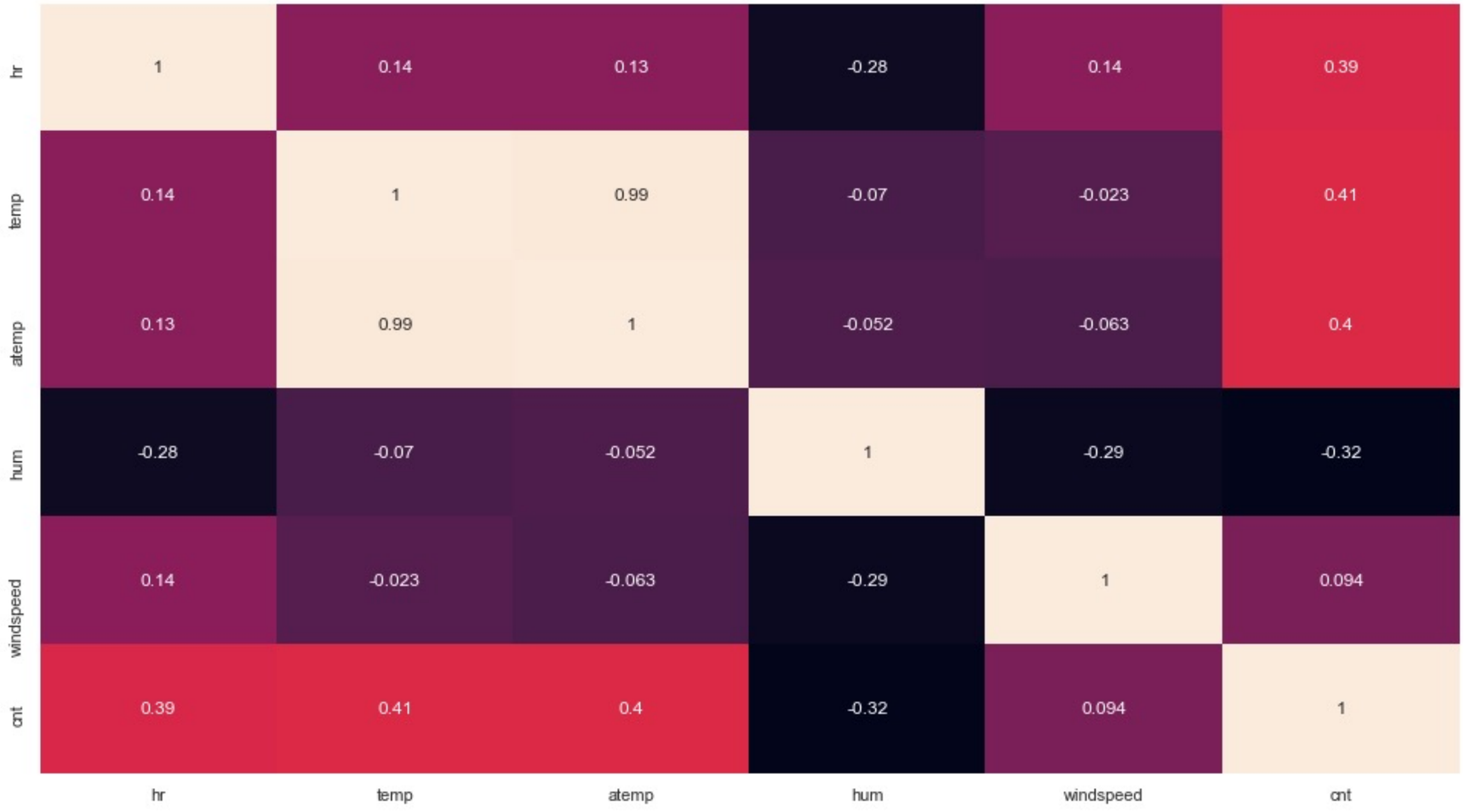
CORRELATION PLOTS FOR DIFFERENT FEATURES IN THE DATASET



HUMIDITY SEEM TO HAVE A NEGATIVE IMPACT WHILE TEMPERATURE AND WINDSPEED HAVE A POSITIVE IMPACT



MOST PEOPLE USE IT WHEN IT IS CLEAR OR PARTLY CLOUDY



NO CORRELATION BETWEEN EXPLANATORY VARIABLES ONLY WITH TARGET VARIABLE

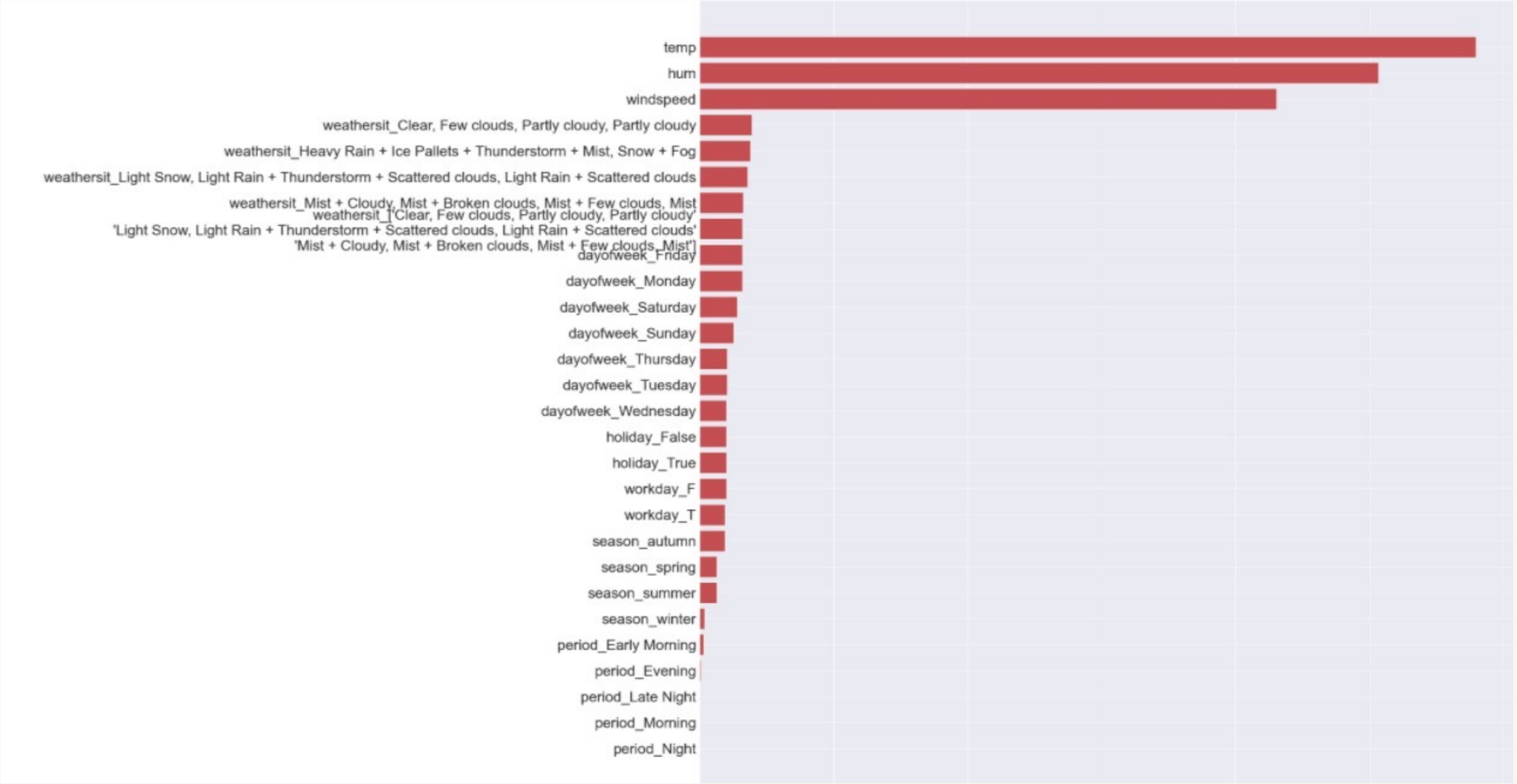
DATA TRANSFORMATION & DATA ENGENEERING

MODIFICATIONS ON THE DATA

- Splitting of the timedate column into Day, Month and Year
- Extraction of US Federal Holidays
- Creation of the different seasons
- Dividing the day into different periods
- Changing the type of certain columns in to categorical and impute missing values with median
- Feature importance and selection based on chi-squared
- Splitting of data into train and test

MODELLING & PARAMETER TUNING

FEATURE IMPORTANCE



MODELLING & PARAMETER TUNING

MODELS

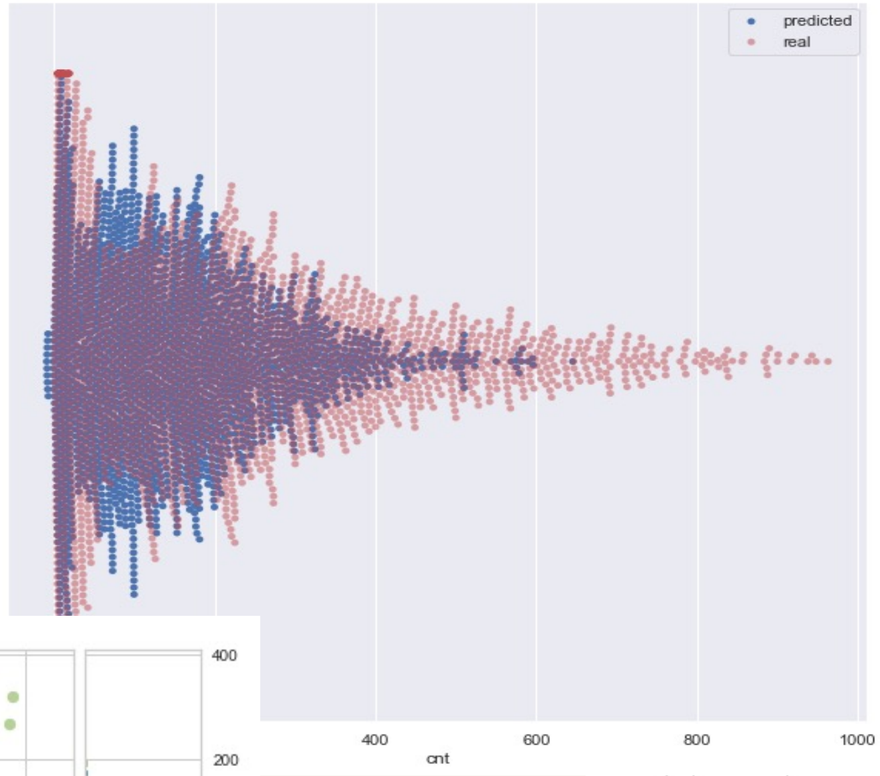
	Model	MAE	MSE	RMSE	R2	RMSLE	MAPE	TT (Sec)
0	Extreme Gradient Boosting	68.3130	10223.3450	101.0688	0.6733	0.8690	1.6566	0.4555
1	Random Forest	67.6815	10358.0253	101.7283	0.6690	0.8243	1.5673	0.5706
2	Linear Regression	88.0528	14584.0316	120.7289	0.5344	1.0811	2.5062	0.0078
3	Ridge Regression	88.0329	14582.1098	120.7211	0.5344	1.0808	2.4987	0.0051
4	Kernel Ridge	88.0334	14582.0566	120.7209	0.5344	1.0808	2.4991	6.2381
5	Lasso Regression	87.8868	14676.7894	121.1134	0.5314	1.0723	2.4467	0.0080
6	Decision Tree	87.3781	18680.8778	136.5907	0.4032	0.9999	1.6682	0.0519

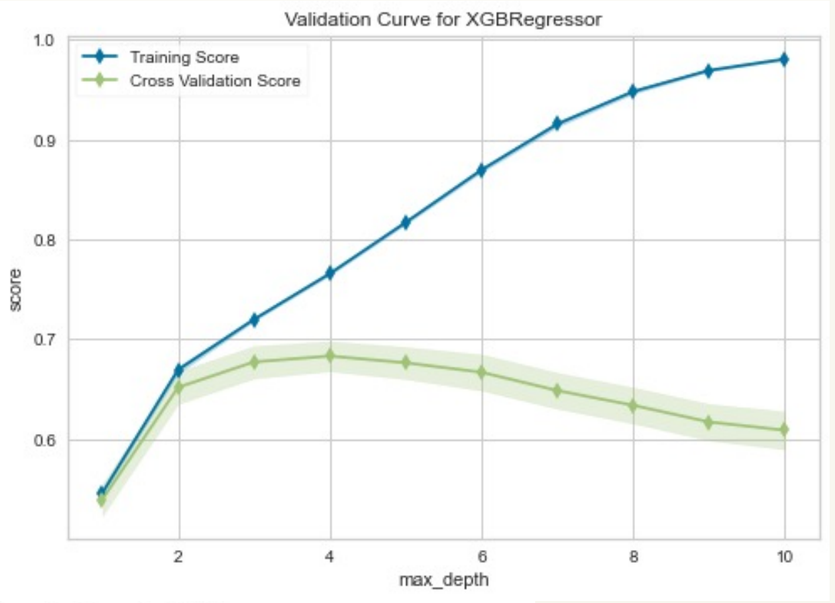
CROSS VALIDATION FOR DIFFERENT MODELS

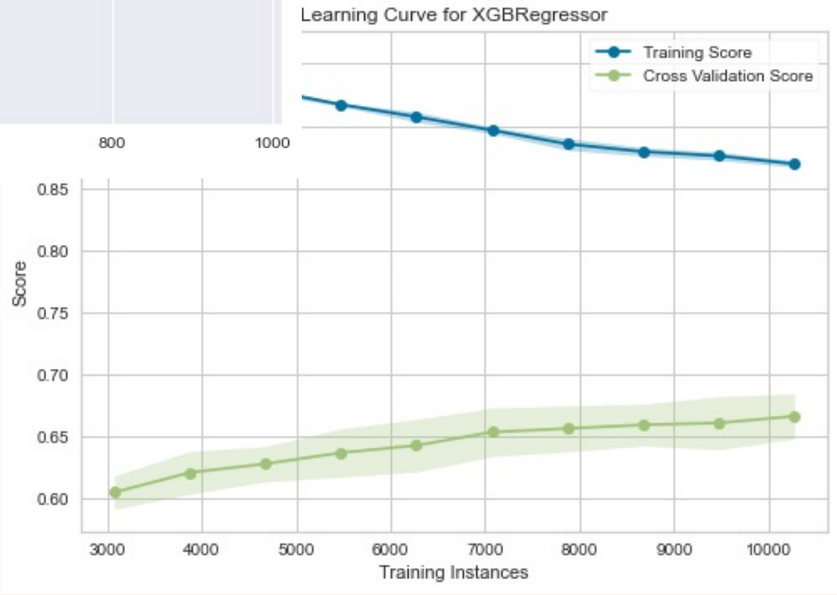
	MAE	MSE	RMSE	R2	RMSLE	MAPE
0	69.7329	9915.2284	99.5752	0.6719	0.8909	2.0793
1	72.1534	10901.9855	104.4126	0.6583	0.8866	2.0811
2	72.4944	10681.9451	103.3535	0.6620	0.8989	2.1173
3	68.8420	9720.4992	98.5926	0.7013	0.8379	1.6946
4	70.6160	10406.5300	102.0124	0.6699	0.8514	1.7888
5	69.2835	10205.9930	101.0247	0.6678	0.8684	1.9635
6	70.4511	10712.9484	103.5034	0.6518	0.8614	1.8745
Mean	70.5105	10363.5899	101.7821	0.6690	0.8708	1.9427
SD	1.2842	406.1220	2.0008	0.0147	0.0207	0.1506

MODELLING & PARAMETER TUNING

INSIGHTS OF THE MODELS







THE TEAM

ALEXANDRE BOUAMAMA

DENNIS PEDERSEN

EUGEN WETTSTEIN

MILAN MEDINA

TONY MATTA

UMUT VAROL

VIMAL KRISHNAN