Отчет о ДЗ №3
Дисциплина: Прикладная мат. статистика
Исполнительница: Смолкина Ю.А
Группа: АДБМ2021

Come up with reasonable goals and methods for statistical processing of data and implement them.

I chose the dataset: https://www.kaggle.com/imakash3011/rental-bike-sharing

Tools: Python (collab), pandas, numpy, matplotlib.pyplot, seaborn, sklearn.model_selection, train_test_split, StandardScaler and more

Initial algorithm and code:

part with ML:

https://colab.research.google.com/drive/13URXYNqdGk5SKxmXRFn900DQH_XTYq2C?usp=sharing

Analysis part:

https://colab.research.google.com/drive/1gnZMqbW0G9emeb7AX77XzlrG7wBmBhQ7?usp=sharing

Objectives: To study the dataset (to find out the distributions of data and their statistical characteristics). Conduct a comparative analysis of the data. Predict the total number of rented bikes in different periods of the day, seasons (perform training and predictions) evaluate the results by metrics (see below)

Metrics: Data distribution, feature correlation, data validity, MSE, R^2, MAPE metrics, L1, 2 regularizations and others

Methods: Data preprocessing, data mining, data validation, machine learning methods, Linear models (Linear regression, Lasso, Ridge and Elastic mesh), Nonlinear models (SVM, random forest, NN) categorical feature normalization, loss functions, RandomForestRegressor, BaseEstimator , TransformerMixin, MannWhitneyU test and others

Introduction: This kind of analysis is very common and useful for young startups, as well as for tourists and people who like to spend time actively. Based on the results of the analysis, a system for online recording and regulation of demand for bicycles in different weather, time of day, season, and in the future even the geographical location of the point of issue of equipment can be developed. Such an analysis was the basis for demand and price regulation in such large companies as Yandex, Uber and others.

Description of the system: Bicycle rental systems are a new generation of traditional bicycle rentals in which the entire process from membership, rental and return has become automatic. With these systems, the user can easily rent a bike at a certain location and return back at another location. Currently, there are about 500 bike-sharing programs in the world, involving more than 500,000 bikes. Today there is a lot of interest in these systems due to their important role in traffic, environment and health issues.

In addition to interesting real world applications of bike sharing systems, the characteristics of the data generated by these systems make them attractive for research. Unlike other transport services such as bus or metro, these systems clearly record the duration of the trip, the place of departure and arrival. This feature turns the bike-sharing system into a virtual sensor network that can be used to detect mobility in the city. Therefore, it is expected that most important events in the city can be detected by monitoring this data.

Source attributes (columns):

```
- instant: record index
- dteday : date
- season : season (1:springer, 2:summer, 3:fall, 4:winter)
- yr : year (0: 2011, 1:2012)
- mnth : month ( 1 to 12)
- hr : hour (0 to 23)
- holiday : weather day is holiday or not (extracted from http://dchr.dc.gov/page/holiday-schedule)
- weekday : day of the week
- workingday : if day is neither weekend nor holiday is 1, otherwise is 0.
+ weathersit :
        - 1: Clear, Few clouds, Partly cloudy, Partly cloudy
        - 2: Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist
        - 3: Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds
        - 4: Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog
- temp : Normalized temperature in Celsius. The values are divided to 41 (max)
- atemp: Normalized feeling temperature in Celsius. The values are divided to 50 (max)
- hum: Normalized humidity. The values are divided to 100 (max)
- windspeed: Normalized wind speed. The values are divided to 67 (max)
- casual: count of casual users
- registered: count of registered users
- cnt: count of total rental bikes including both casual and registered
```

Let's look at our data:

	instant	dteday	season	yr	mnth	hr	holiday	weekday	workingday	weathersit	temp	atemp	hum	windspeed	casual	registered	cnt
0	1	2011-01-01	1	0	1	0	0	6	0	1	0.24	0.2879	0.81	0.0000	3	13	16
1	2	2011-01-01	1	0	1	1	0	6	0	1	0.22	0.2727	0.80	0.0000	8	32	40
2	3	2011-01-01	1	0	1	2	0	6	0	1	0.22	0.2727	0.80	0.0000	5	27	32
3	4	2011-01-01	1	0	1	3	0	6	0	1	0.24	0.2879	0.75	0.0000	3	10	13
4	5	2011-01-01	1	0	1	4	0	6	0	1	0.24	0.2879	0.75	0.0000	0	1	1
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
17374	17375	2012-12-31	1	1	12	19	0	1	1	2	0.26	0.2576	0.60	0.1642	11	108	119
17375	17376	2012-12-31	1	1	12	20	0	1	1	2	0.26	0.2576	0.60	0.1642	8	81	89
17376	17377	2012-12-31	1	1	12	21	0	1	1	1	0.26	0.2576	0.60	0.1642	7	83	90
17377	17378	2012-12-31	1	1	12	22	0	1	1	1	0.26	0.2727	0.56	0.1343	13	48	61
17378	17379	2012-12-31	1	1	12	23	0	1	1	1	0.26	0.2727	0.65	0.1343	12	37	49

17379 rows × 17 columns

Исходный размер: 17379 rows × 17 columns

I. Data Preprocessing

Any work and subsequent data analysis should begin with preprocessing. We have to study outliers that can introduce noise into our data, forecasts. When building hypotheses and from refutation / confirmation, it is better to rely on cleaned data.

Remove the instant column (This column corresponds to the observation number and does not make sense)

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 17379 entries, 0 to 17378
Data columns (total 16 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   dteday      17379 non-null  object
 1   season      17379 non-null  int64
 2   yr          17379 non-null  int64
 3   mnth        17379 non-null  int64
 4   hr          17379 non-null  int64
 5   holiday     17379 non-null  int64
 6   weekday     17379 non-null  int64
 7   workingday  17379 non-null  int64
 8   weathersit  17379 non-null  int64
 9   temp        17379 non-null  float64
 10  atemp       17379 non-null  float64
 11  hum         17379 non-null  float64
 12  windspeed   17379 non-null  float64
 13  casual      17379 non-null  int64
 14  registered  17379 non-null  int64
 15  cnt         17379 non-null  int64
dtypes: float64(4), int64(11), object(1)
memory usage: 2.1+ MB
```

And let's look at the statistical values (scatter of data) in each column

	season	yr	mnth	hr	holiday	weekday	workingday	weathersit	temp	atemp
count	17379.000000	17379.000000	17379.000000	17379.000000	17379.000000	17379.000000	17379.000000	17379.000000	17379.000000	17379.000000
mean	2.501640	0.502561	6.537775	11.546752	0.028770	3.003683	0.682721	1.425283	0.496987	0.475775
std	1.106918	0.500008	3.438776	6.914405	0.167165	2.005771	0.465431	0.639357	0.192556	0.171850
min	1.000000	0.000000	1.000000	0.000000	0.000000	0.000000	0.000000	1.000000	0.020000	0.000000
25%	2.000000	0.000000	4.000000	6.000000	0.000000	1.000000	0.000000	1.000000	0.340000	0.333300
50%	3.000000	1.000000	7.000000	12.000000	0.000000	3.000000	1.000000	1.000000	0.500000	0.484800
75%	3.000000	1.000000	10.000000	18.000000	0.000000	5.000000	1.000000	2.000000	0.660000	0.621200
max	4.000000	1.000000	12.000000	23.000000	1.000000	6.000000	1.000000	4.000000	1.000000	1.000000

*Not all columns shown

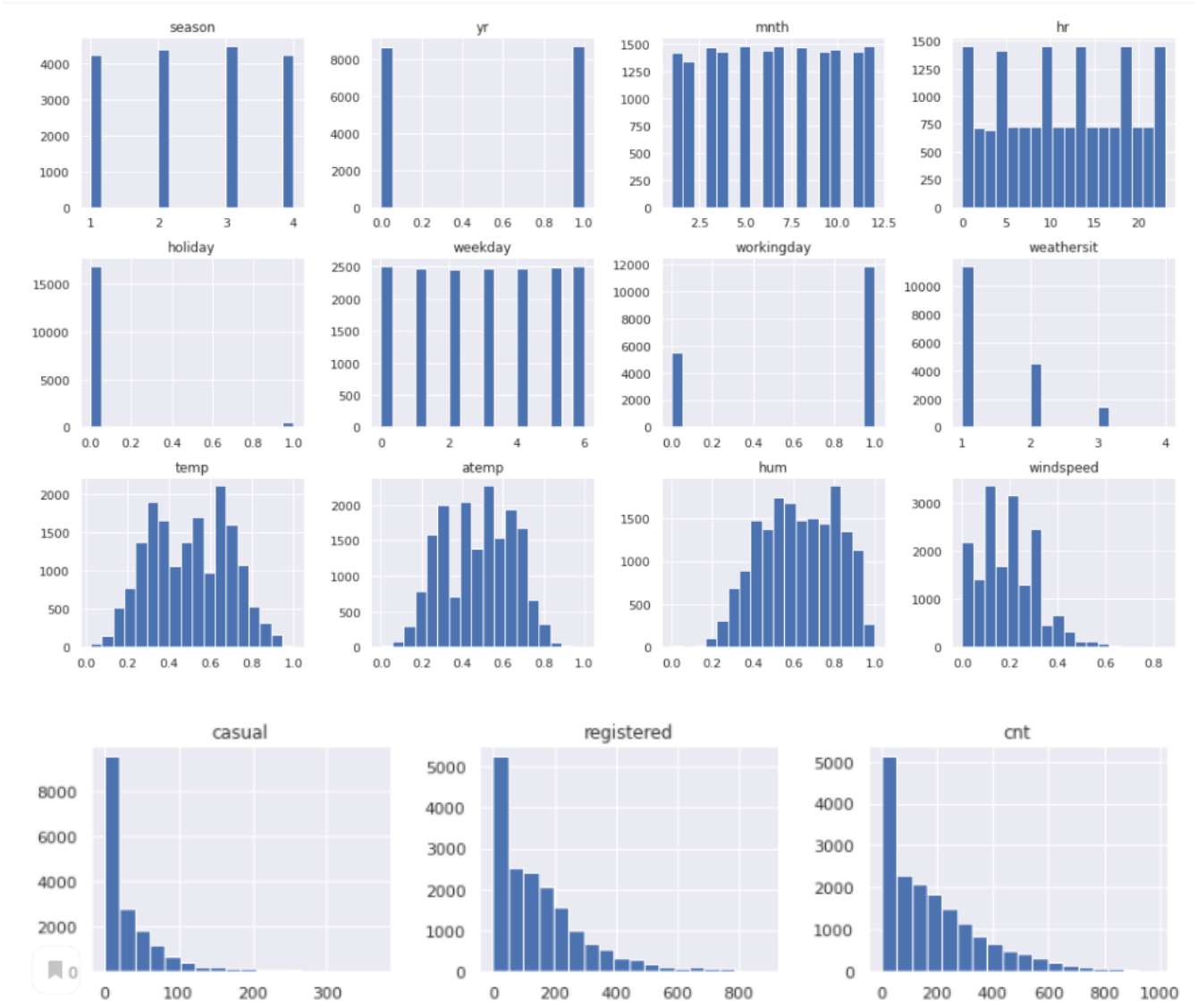
Note right away that the Minimum value for rented bikes is 1. It can be assumed that rental information was not recorded if the bikes were not rented.

Let's define how the data is distributed and how it is correlated:

We can make multiple charts to visualize your data.

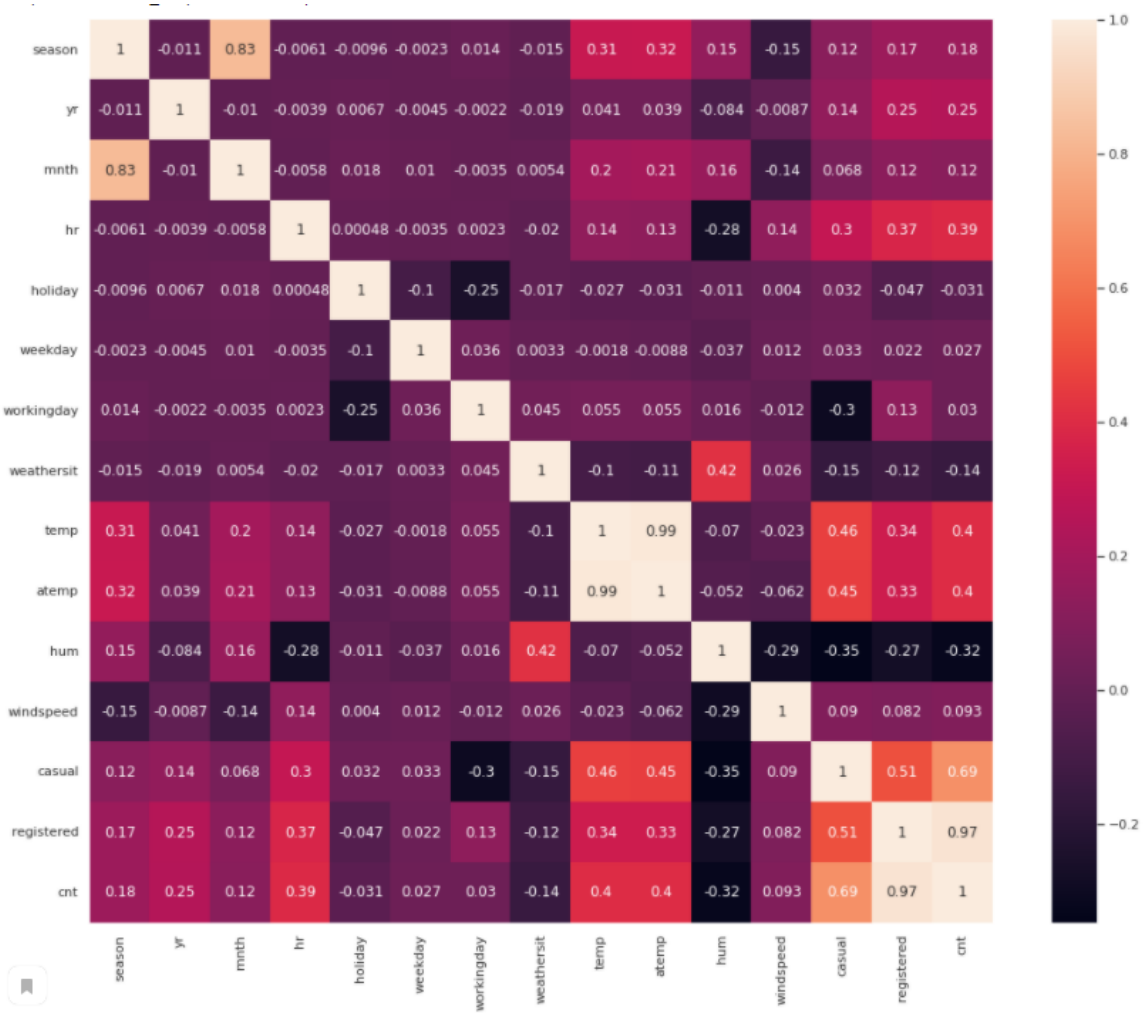
Histograms to describe the distribution of each feature

Heat map matrix to estimate the correlation between each pair of columns

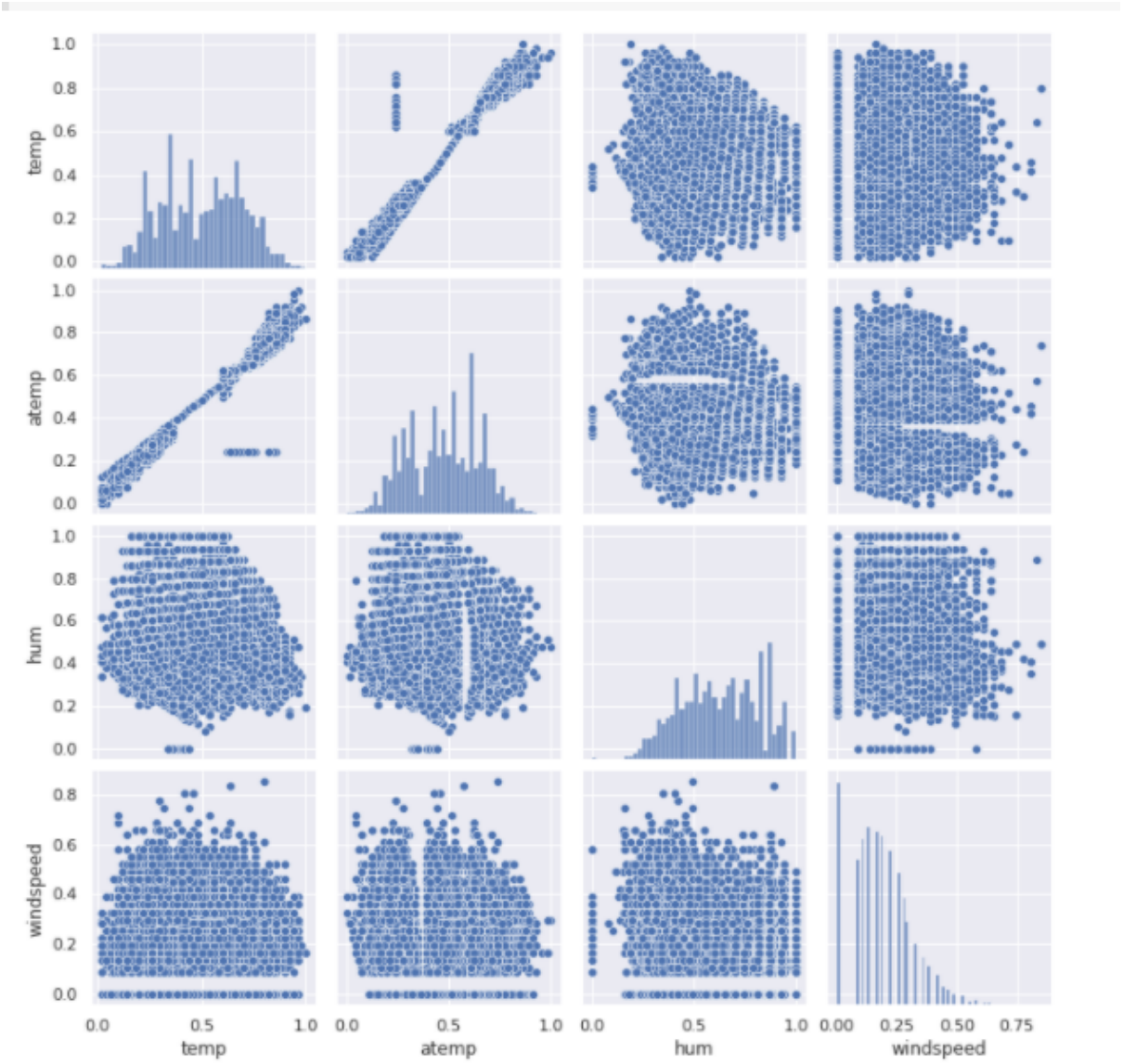


- Looking at these histograms, we can make some points:
- The features 'season', 'mnth', 'yr', 'dayday' have a discrete uniform distribution, which means that our data is balanced "in time". The data contains almost the same number of records for different seasons, years, days of the week.
- The distribution of the feature 'hr' does not look uniform, which may be due to the assumption made earlier: if no bicycles are rented at this hour, then information is not recorded.
- The information about the working day looks a little strange, because. working days in the usual sense is 5 days out of seven, but the ratio of the heights of the bars is more like a 1:2 ratio. Holidays are also considered non-working days, but on the "Holidays" histogram, we see that there are not many such days.
- The distributions of the signs temp, atemp and hum (humidity) are close to the norm; polymodality can be seen in the distributions of the characteristics "temp" and "atemp"; positive asymmetry can be observed in the distribution of signs of "wind speed".
- The target feature "cnt" seems to have an exponential distribution (as well as "random" and "registered") or could be a Poisson distribution with $\lambda = 1$.
-

Feature correlation matrix:



- As we can see, the functions "temp" and "atemp" are highly correlated. This is expected because temp is the normalized temperature in degrees Celsius and atemp is the normalized feel temperature in degrees Celsius.
- A high correlation between "season" and "month" is also evident. Information about both functions may be redundant.
- 'cnt' is the sum of 'random' and 'registered', so this feature is related to both of them; but it correlates more with "registered". The reason may be that there are more registered users than random ones, and they contribute the most to the total amount and determine its behavior. Next, we will dwell on this issue a little.



As we can see, there is a dependency between the temp and atemp functions, so we can omit one of them. We can also see a strange peak at "zero" for the "wind speed" function, so we can try to exclude these observations from the sample. But there are many observations with a value of 0 on this feature, and they do not look like outliers, they probably reflect completely calm weather.

Let's depict how the average monthly number of bikes for rent changes during the year. Let's draw curves for both 2011 and 2012.

As you can see, for each month the average monthly number of rented bikes in 2012 is higher than in 2011. This may be due to the growing popularity of bicycles, with the improvement of the rental system, and so on. It can be concluded that the trait is important because there is a difference of two years. The year column should not be removed from the feature matrix.

However, it should be noted that there are some nuances to the real problem: in some years, the number of users may reach a plateau and stop growing from year to year, or, conversely, the use of bike sharing may fall. for some reason (isolation due to coronavirus infection, for example). But in our case, we use data only for these two years and will not go into such nuances.

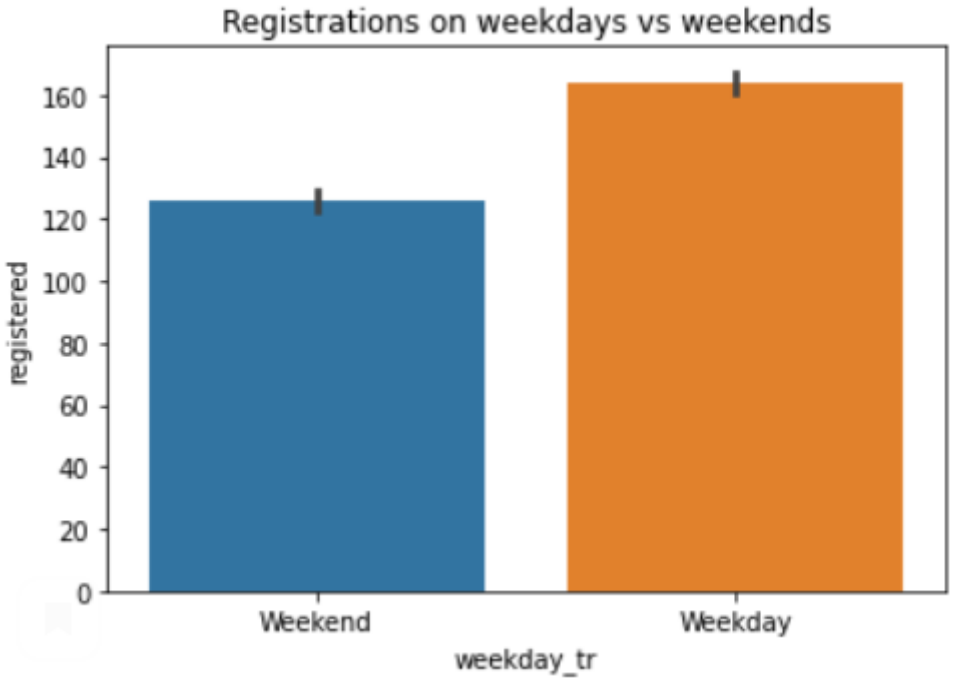
Task: As described at the very beginning, I want to analyze the number of registrations (issues of bicycles) depending on various characteristics: weather, season, day of the week (weekday or weekend), etc.

Ideas:

- Divide the time of the day into 5 categories
- Early morning - from 4 to 7 o'clock
- Morning - from 8 to 12 hours.
- Second half of the day - from 13 to 16 hours.
- Evening - from 17 to 20 hours.
- Night - from 21 o'clock to 3 o'clock

1. Divide days into weekdays or weekends
2. Season, Weather Sit - change it to OHE as it is not ordinal data.

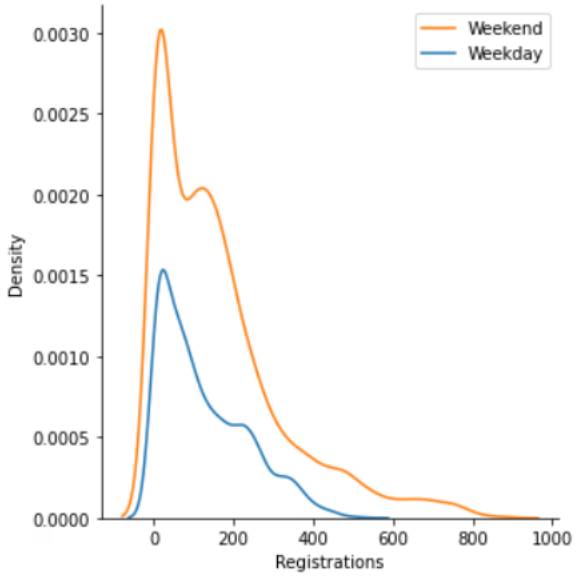
We find that Weekends are not considered holidays. However, in two years, 412 days off.



Conclusion: Surprisingly, on weekdays there are more registrations and releases of bicycles than on weekends. One reason could be that people use bikes for their daily commute to work or school.

We will use the MannWhitneyU test to see if the number of signups varies by day of the week or weekend.

Even though the MannWhitneyU test does not assume that the dependent variable is normally distributed, let's go ahead and see if it is normally distributed.



Checking for value skewness

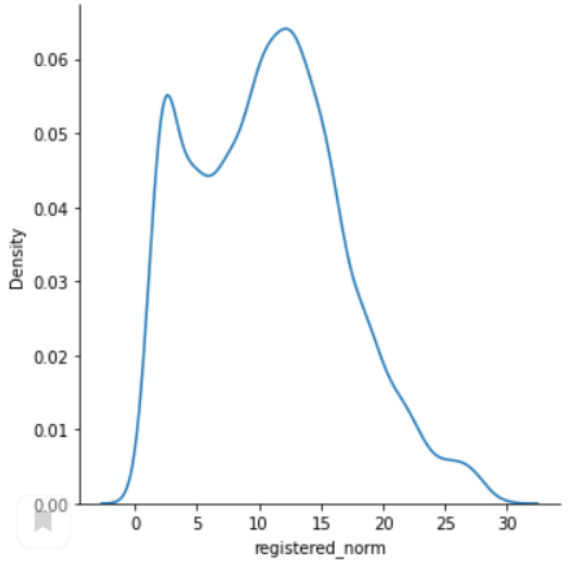
H0 - This function tests the null hypothesis that the skewness of the population from which the sample was taken is the same as that of the corresponding normal distribution.

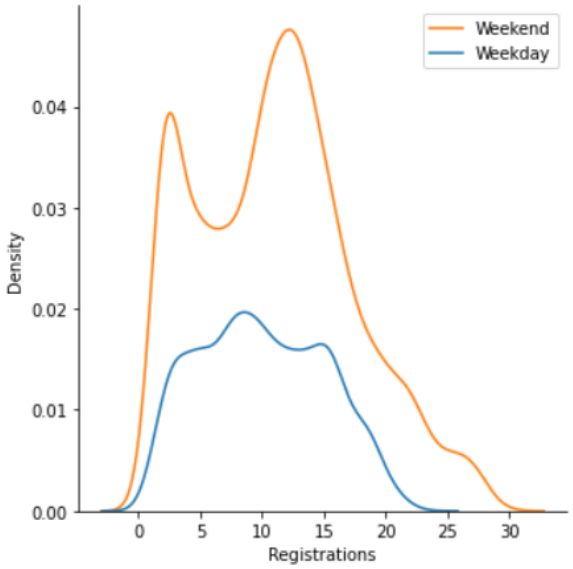
```
temp : 0.6706915425920954 ['Normal']
atemp : 0.00022011874108668952 ['Skewed']
registered : 0.0 ['Skewed']
casual : 0.0 ['Skewed']
**********************************************
temp : 0.008828526626182949
atemp : -0.07682598398053479
registered : 1.5527844974336837
casual : 2.538369552597277
```

It can be seen that the “Recorded Quantity” is not normally distributed, which is confirmed by skewtest

After some transformation of the data representation:

	registered	registered_norm
8273	235	15.329710
8393	220	14.832397
1351	61	7.810250
12541	726	26.944387
12862	4	2.000000
...	...	...
9475	54	7.348469
3714	199	14.106736
7409	418	20.445048
5787	176	13.266499
14155	99	9.949874





Although the dependent variable is not normally distributed, the Mann-Whitney test does not require the target to be normally distributed.

Hypothesis testing whether the number of bookings change depending on the day of the week

The Mann-Whitney U test is a nonparametric test of the null hypothesis that the distribution underlying sample x is the same as the distribution underlying sample y. It is often used to check for location differences between distributions.

p-value of test is 0.0000. Hence, the null hypo is rejected.

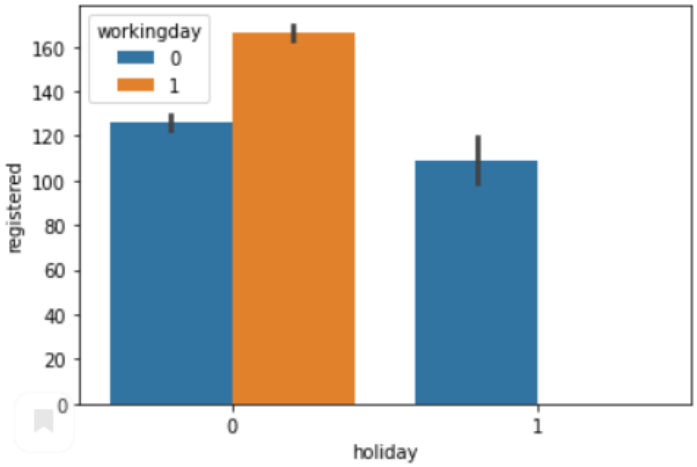
Therefore, there is indeed a difference in registered - used equipment between weekdays and weekends.

The result seems intuitive and logical.

Let's now look at weekends in more detail.

Holidays vs Registrations

Assumption: There will be more bookings over the weekend



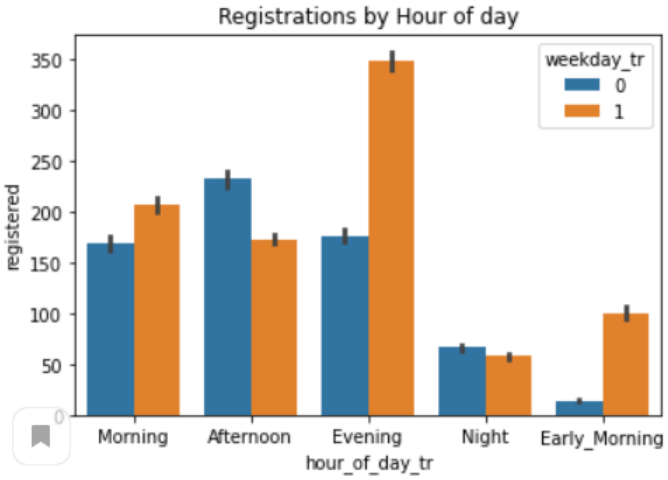

```
holiday  workingday  weekday_tr
0        0           0            4027
         1           1            9464
1        0           1             412
Name: registered, dtype: int64
```

There is more demand on non-holiday days. This also indicates that people mainly use bicycles for daily commuting to work, study, and personal matters. The weekday/weekend relationship with “check-ins” also confirmed this.

It would also be interesting to break the day into temporary shorter periods, let's do it

Registrations by Hour of Day

Assumption: There will be more bookings in the morning



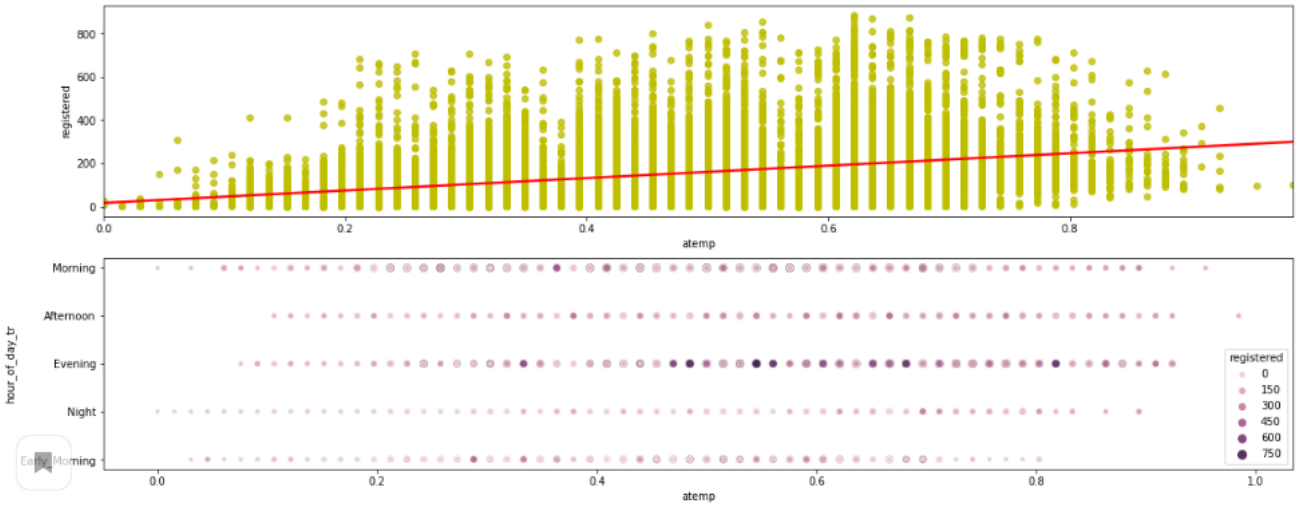
Conclusions:

- There is more demand in the evening than during the day.
-
- There are fewer overnight bookings than at other times on weekdays. However, there are fewer morning bookings on weekends.
-
- Weekday evenings are more popular than mornings. People can commute to their homes/nearby areas (from work) in the evenings.
-
- Weekend afternoons also register more demand than any other weekend.

I would suggest raising prices for booking equipment on weekday evenings, weekend afternoons.

Assumption: Also since cycling is an open mode of transport, so the temperature, season will have a big impact on the number of bookings.

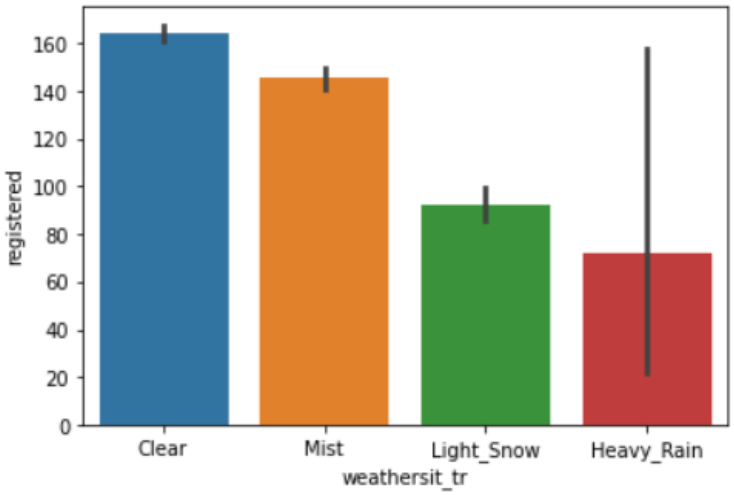
Registrations by temperature



We will try to confirm this dependence by checking the correlation coefficient for continuous variables

The regular chart shows that the number of registrations will increase as temperatures rise.

Погода:



	registered		
	min	max	mean
weathersit_tr			
Clear	0	886	164.098229
Heavy_Rain	22	158	71.666667
Light_Snow	0	734	92.492091
Mist	0	822	145.327976

When the weather is clear or foggy, demand is greater. However, it is very unpredictable when it rains.

After data encoding, the following attributes remained:

```
['yr',
 'mnth',
 'holiday',
 'workingday',
 'temp',
 'atemp',
 'hum',
 'windspeed',
 'casual',
 'registered',
 'cnt',
 'weekday_tr',
 'registered_norm',
 'season_tr_spring',
 'season_tr_summer',
 'season_tr_winter',
 'weathersit_tr_Heavy_Rain',
 'weathersit_tr_Light_Snow',
 'weathersit_tr_Mist',
 'hour_of_day_tr_Early_Morning',
 'hour_of_day_tr_Evening',
 'hour_of_day_tr_Morning',
 'hour_of_day_tr_Night']
```

Name for each categorical attribute MannWhitneyU test p-values

```
p-value for field :holiday is 0.000, null hyp rejected
p-value for field :workingday is 0.000, null hyp rejected
p-value for field :weekday_tr is 0.000, null hyp rejected
p-value for field :season_tr_spring is 0.000, null hyp rejected
p-value for field :season_tr_summer is 0.000, null hyp rejected
p-value for field :season_tr_winter is 0.000, null hyp rejected
p-value for field :weathersit_tr_Heavy_Rain is 0.192, null hyp not rejected
p-value for field :weathersit_tr_Light_Snow is 0.000, null hyp rejected
p-value for field :weathersit_tr_Mist is 0.004, null hyp rejected
p-value for field :hour_of_day_tr_Early_Morning is 0.000, null hyp rejected
p-value for field :hour_of_day_tr_Evening is 0.000, null hyp rejected
p-value for field :hour_of_day_tr_Morning is 0.000, null hyp rejected
p-value for field :hour_of_day_tr_Night is 0.000, null hyp rejected
```

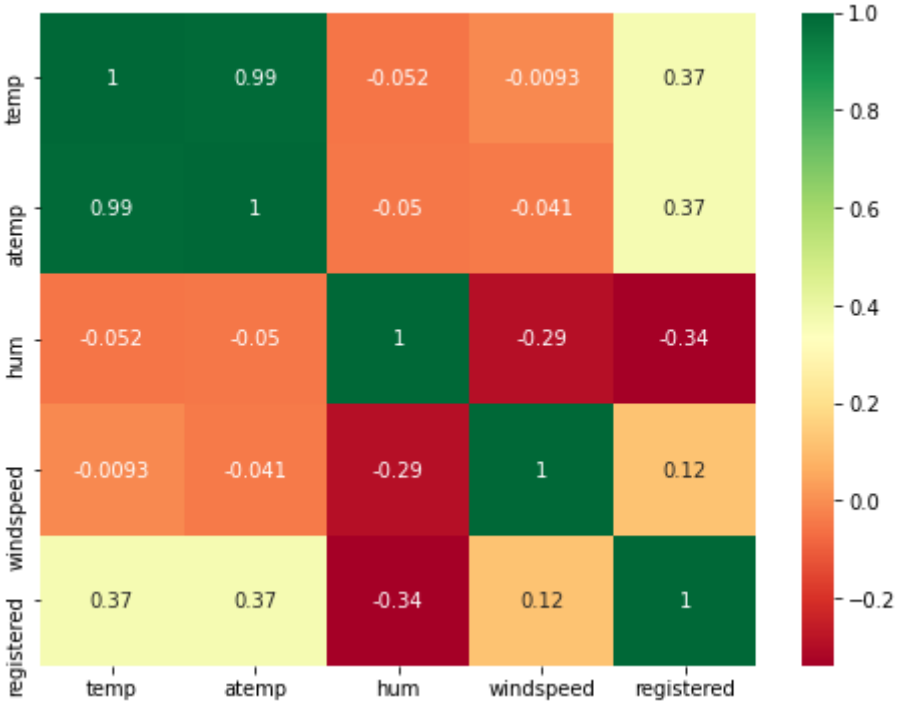
The default or null hypothesis is that there is no difference between the distributions of the data samples. Rejection of this hypothesis suggests that there is probably some difference between the samples.

All categorical values, with the exception of HARD rain, appear to affect demand.

	temp	atemp	hum	windspeed	registered
8273	0.36	0.3333	0.46	0.3881	235
8393	0.40	0.4091	0.87	0.2239	220
1351	0.32	0.3333	0.36	0.1343	61
12541	0.80	0.7273	0.43	0.2985	726
12862	0.68	0.6364	0.69	0.2537	4
...	...	...	...	...	...
9475	0.30	0.2879	0.70	0.1940	54
3714	0.92	0.8788	0.40	0.2239	199
7409	0.40	0.4091	0.94	0.0896	418
5787	0.72	0.6818	0.66	0.1642	176
14155	0.64	0.5758	0.83	0.2239	99

13903 rows × 5 columns

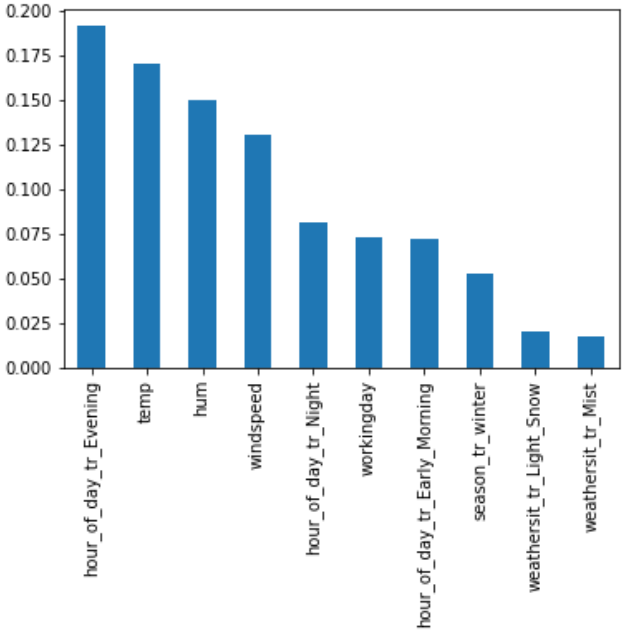
Look at the correlation matrix, I'll use the 'spearman' method, since it's not a normal distribution.



Only 2 columns i.e. temperature and humidity are moderately correlated with demand.

For the final check, I weighted the importance of the parameters using PermutationImportance, RandomForestRegressor

Weight	Feature
0.3529 ± 0.0422	hour_of_day_tr_Evening
0.2573 ± 0.0190	hour_of_day_tr_Night
0.2467 ± 0.0168	temp
0.1577 ± 0.0213	workingday
0.1573 ± 0.0190	hour_of_day_tr_Early_Morning
0.1376 ± 0.0110	hum
0.0934 ± 0.0173	season_tr_winter
0.0319 ± 0.0115	season_tr_summer
0.0285 ± 0.0044	weathersit_tr_Light_Snow
0.0260 ± 0.0089	windspeed
0.0226 ± 0.0026	hour_of_day_tr_Morning
0.0155 ± 0.0087	weekday_tr
0.0098 ± 0.0053	weathersit_tr_Mist
0.0073 ± 0.0078	season_tr_spring
0.0021 ± 0.0004	holiday
0 ± 0.0000	weathersit_tr_Heavy_Rain

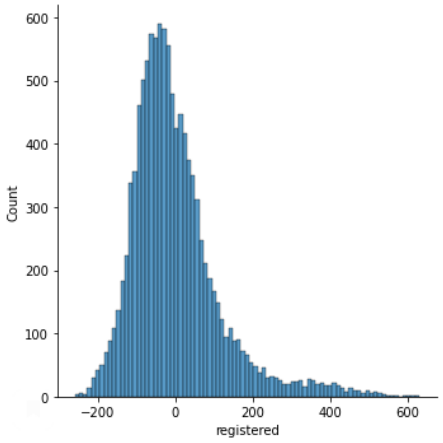


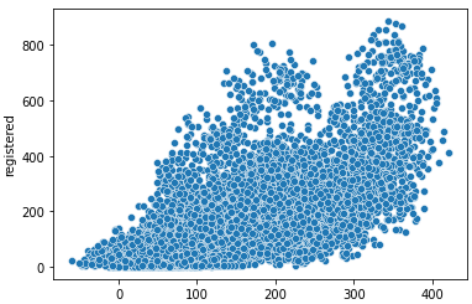
The prediction showed that

Linear regression scores do not differ much between training and testing sets.

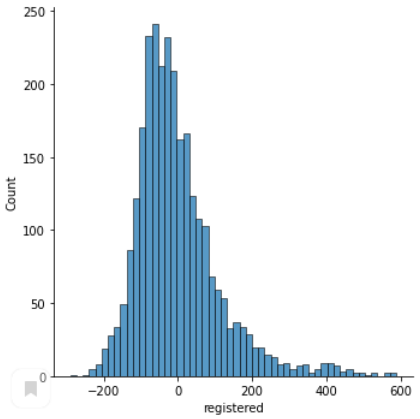
Education:

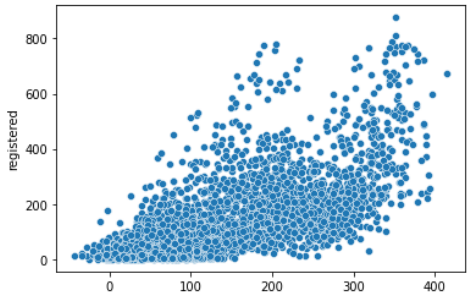
In general, it can be seen that the errors are normally distributed





Validation:





Erros comparison:

Training error : 115.172

Validation error : 113.472

Other methods are in the jupytar for ML part

Summary:

After feature selection, the following variables were found to be important compared to others.

- hour_of_day_tr_Evening
- temp
- hour_of_day_tr_Night
- hour_of_day_tr_Early_Morning
- hum
- workingday
- windspeed

Although the linear regressor was stable between the training and test sets, the random forest regressor predicts with the smallest error. The score improved even more after hyperparameter tuning using Randomized Search CV and the best score was used to validate the score/performance of the test data.

Explanations:

The neural network model showed better results than the models of other algorithms. It has the largest coefficient of determination R2 on the test set and the smallest RMSE. Best settings: Learning_rate=0.05 and epochs=80. The NN performance in the training set is better than in the test, but not significantly, and it is normal that the performance of the model on the training data is higher.

The SVR model also performed very well and showed the second result in terms of metrics on the test set. Best parameters: cost value C=1000 and kernel=rbf.

The Gradient Boosting model comes in 3rd with hyperparameters: max depth = 8, min child_mass = 4, and gamma = 0.4.

Although the Random Forest model has the best performance on the training data (R2 = 0.987!), but on the test data it takes only 4th place. It can be assumed that in this case there was a slight refitting of the model. However, the model still performs well on the test data, so I can't be sure if overfitting is taking place.

Temperature and humidity are the most important characteristics in the results of the Gradient Boosting and Random Forest models.

Linear models (Linear Regression, Lasso, Ridge and Elastic Mesh) have the lowest R2 score (more than 0.2 worse than the models listed in the previous paragraphs) and the RMSE score for them is about 2 times greater. Perhaps this is due to the fact that the models are too simple to describe patterns in the data. In addition, there are a lot of categorical features in the data, while there are practically no real features. All these models give almost the same result and no significant improvements can be found. Various types of regularization are designed to help combat overfitting and multicollinearity, in our case it was rather underfitting. Pairwise correlated features were found at the pre-processing stage, and one feature in a pair was discarded before the models were tuned.

However, prediction errors can range from 100 to 109.

Further actions. A more accurate result can be obtained if GridSearchCV is used instead of RandomizedSearchCV when tuning hyperparameters and if other models can be explored. A similar model should be created for "random" counts, and then "registered" and "random" counts can be summed to determine "Cnt", that is, the total count.

Also, for practical use, I recommend paying attention to the midterm elections on the dependence of the number of registrations depending on the weather, time.

```
best_results
```

	R2_train	RMSE_train	R2_test	RMSE_test	Best parameters
Linear regression	0.702	99.153	0.689	100.245	-
Ridge regression	0.702	99.155	0.689	100.243	alpha_l = 1.024
Lasso	0.702	99.162	0.689	100.248	alpha_ll = 0.006
ElasticNet	0.702	99.167	0.689	100.253	alpha = 0.01, l1_ratio = 0.99
SVM	0.967	33.204	0.931	47.266	C = 1000, kernel = rbf
Random Forest	0.987	20.701	0.911	53.683	n_estimators = 95, max_features = 42
Gradient Boosting	0.947	41.934	0.917	51.950	max_depth = 8, min_child_weight = 4, gamma = 0.4
Neural Network	0.971	30.787	0.941	43.530	epochs = 80, learning_rate = 0.05

Demand Analysis

- Demand is higher on weekdays than on weekends.
- There is a difference in registered - used equipment between weekdays and weekends.
- There is more demand on non-holiday days.
- There is more demand in the evening than during the day.

- There are fewer overnight bookings than at other times on weekdays. However, there are fewer morning bookings on weekends.
- Weekday evenings are more popular than mornings. People can commute to their homes/nearby areas (from work) in the evenings.
- Weekend afternoons also register more demand than any other weekend.
- When the weather is clear or foggy, demand is greater. However, it is very unpredictable when it rains.
- All categorical values, with the exception of HARD rain, appear to affect demand.
- Temperature and humidity are moderately correlated with demand.