ANSWER SHEET
DAC 2021

DATA ANALYSIS
COMPETITION
2021

TEAM NAME	Hmm Okay
TEAM ID	ID-21-0114
UNIVERSITY	Universitas Gadjah Mada

a. This dataset consists of many variables as written in the question sheet. To make predictions, we must choose which variables have a strong relevance to the target variable. The method we use to see the relevance between the variables used for prediction and the target variable is pearson correlation and F test regression.

Based on the results of the correlation test, we chose variable price_share_usd, market_share_usd, watchers, stars, stars, contributors_all time, market_cap_rank, company_code, copy because the correlation value is around 0.2 which indicates the correlation is quite strong.

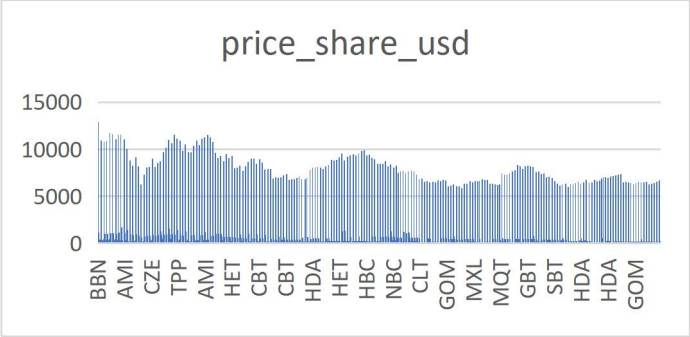
price_share_usd	1.000000
market_cap_usd	0.926513
watchers	0.824209
stars	0.777599
contributors_all_time	0.509665
market_cap_rank	0.361818
company_code	0.261974
copy	0.195527

Then we did the F test. By stating the initial hypothesis that the regression model was not feasible to use and the alternative hypothesis the regression model was feasible to use. At the significance level of =5%, the initial hypothesis will enter the criticism/rejection area if the value is less than the significance level.
After the test we get the F values is less than 5% then the initial hypothesis is rejected. So it can be concluded that the regression model is feasible to use.

```
market_cap_usd            0.000000e+00
watchers                  0.000000e+00
stars                     0.000000e+00
contributors_all_time     0.000000e+00
market_cap_rank          9.230337e-169
company_code_enc          1.737628e-86
copy                      3.027476e-48
```

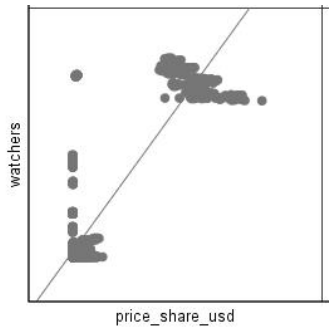
Exploratory Data Analysis

- **Company_code and Price_share_usd**



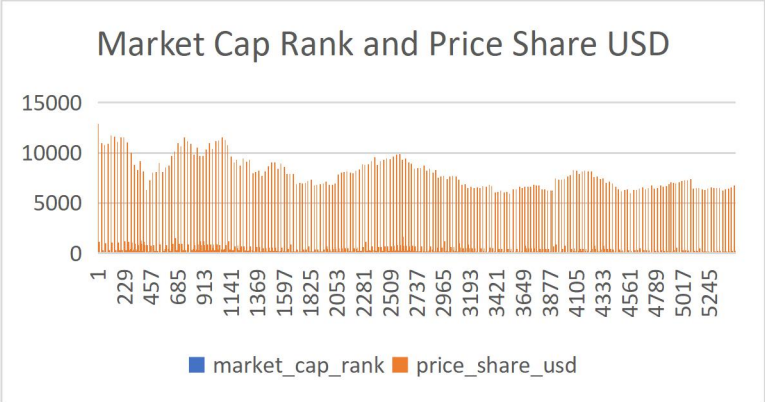
In general, BBN has the highest share value among other companies. The differerence BBN's price share with other companies isn't significantly different. The company "BBN" is a famous worldwide company. The popularity of the company can make the share price rise because many people believe in the sustainability of the company in a situation like this.

- **Watchers and Price_share_usd**



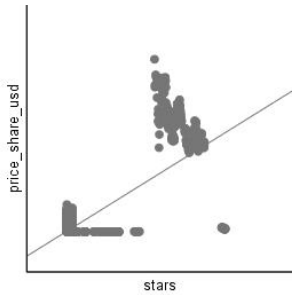
Market sentiment in stocks is important. The plot shows that watchers and price share are positively correlated. If the market sentiment is positive, of course the price share will rise. However, if the market sentiment drops, the share price will also fall.

- **Market_cap_rank**



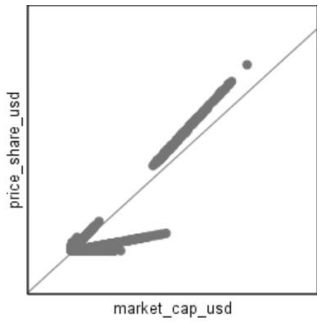
The better the ranking of the company, the stock price will rise. People will think that the management of the company is very good.

- **Stars and Price_share_usd**



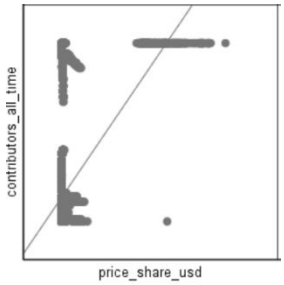
On the plot we can see that the plot is positively correlated. The more people who think that code analysis useful, the price share will go up. If the information provided in the analysis is negative, the price will go down. People who buy shares will think if the analysis is not useful it will lead to more losses if one day the company cannot survive in this economic situation.

- **Market_cap_usd and Price_share_usd**



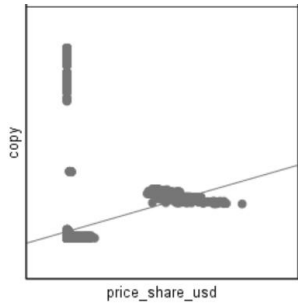
If the value of the company increases, the share price will also increase.

- **Contributor_all_time and Price_share_usd**



The more people analyze a share price movement at all time, the company's share price will increase because it is always updated. People believe that the company can thrive in any situation.

- **Copy and Price_share_usd**



The more people who copy the stock analysis code, the share price will increase. People would think that the company analytics code he copied had a good reputation and would continue to grow.

b. In the next stage, we will create a model and validate the model. We know that the dependent variable in this case is “price_share_usd” which is numeric data. After doing exploratory, we find that there is a significant relationship between the dependent variable "price_share_usd" and the selected independent variable "market_cap_usd","watchers","stars","contributors_all_time","market_cap_rank","company_code_enc","copy".

Because the target variable is numerical data and shows a significant relationship between the selected independent variables, we can use the regression analysis method. In order to reduce RMSE we use a more advanced method, namely XGBoostRegressor.

Gradient boosting refers to a class of ensemble machine learning algorithms that can be used for classification or regression predictive modeling problems.

Ensembles are constructed from decision tree models. Trees are added one at a time to the ensemble and fit to correct the prediction errors made by prior models. This is a type of ensemble machine learning model referred to as boosting.

Models are fit using any arbitrary differentiable loss function and gradient descent optimization algorithm. This gives the technique its name, “gradient boosting,” as the loss gradient is minimized as the model is fit, much like a neural network.

XGboostRegressor is designed to be both computationally efficient and highly effective, perhaps more effective than other open-source implementations.

In XGboostRegressor there are parameters that must be specified, to determine the best parameters, we do tuning. The parameter is

- **booster:** Select the type of model to run at each iteration
  gbtree: tree-based models
  gblinear: linear models
- **nthread:** default to maximum number of threads available if not set

- **objective:** This defines the loss function to be minimized

Parameters for controlling speed

- **subsample:** Denotes the fraction of observations to be randomly samples for each tree.
- **colsample_bytree:** Subsample ratio of columns when constructing each tree.
- **n_estimators:** Number of trees to fit.

Important parameters which control overfiting

- **learning_rate:** Makes the model more robust by shrinking the weights on each step
- **max_depth:** The maximum depth of a tree.
- **min_child_weight:** Defines the minimum sum of weights of all observations required in a child.

We get :
{'colsample_bytree': 0.7,
'learning_rate': 0.1,
'max_depth': 7,
'min_child_weight': 1,
'n_estimators': 500,
'objective': 'reg:squarederror',
'subsample': 0.5}

After getting the best parameters, then testing is carried out. Test data comes from 20% of available datasets.
rsme: 26.709546400476857
mae: 2.6881794746248

Next we validate using Kfold validation
Obtained
mean_RMSE : 34.843187868730332
mean_MAE : 7.9674863334879721

c. Based on our analysis, we conclude that variables market_share_usd, watchers, stars, stars, contributors_all time, market_cap_rank, company_code, copy correlated with share_price_usd.

Internet activity is closely related to share prices. In an economic situation like this, the way for a company to survive is to improve the quality of its share price analysis and also consider variables related to share prices.

The increase or decrease in the company's condition greatly affects the market sentiment where if the company's condition increases directly, the market sentiment will be positive. On the other hand, if the condition of the company declines, the market sentiment will be negative and affect the share price.