# CS5228 PROJECT

GROUP 17


CHAN JUN WEI

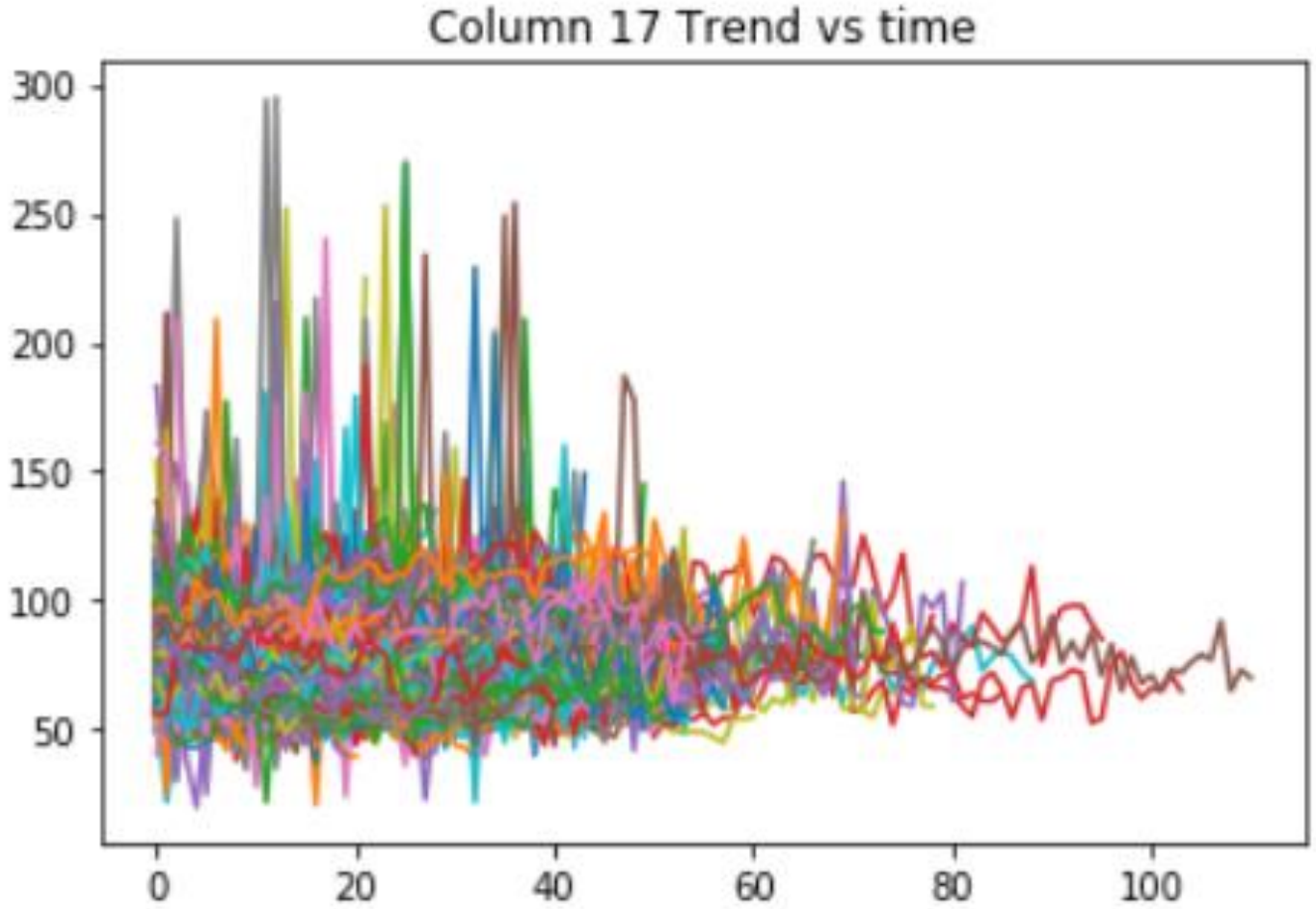
CHUA CHIN SIANG

TEY SHIWEI

WANG ZIHAO

# DATA EXPLORATION

## TREND VS TIME

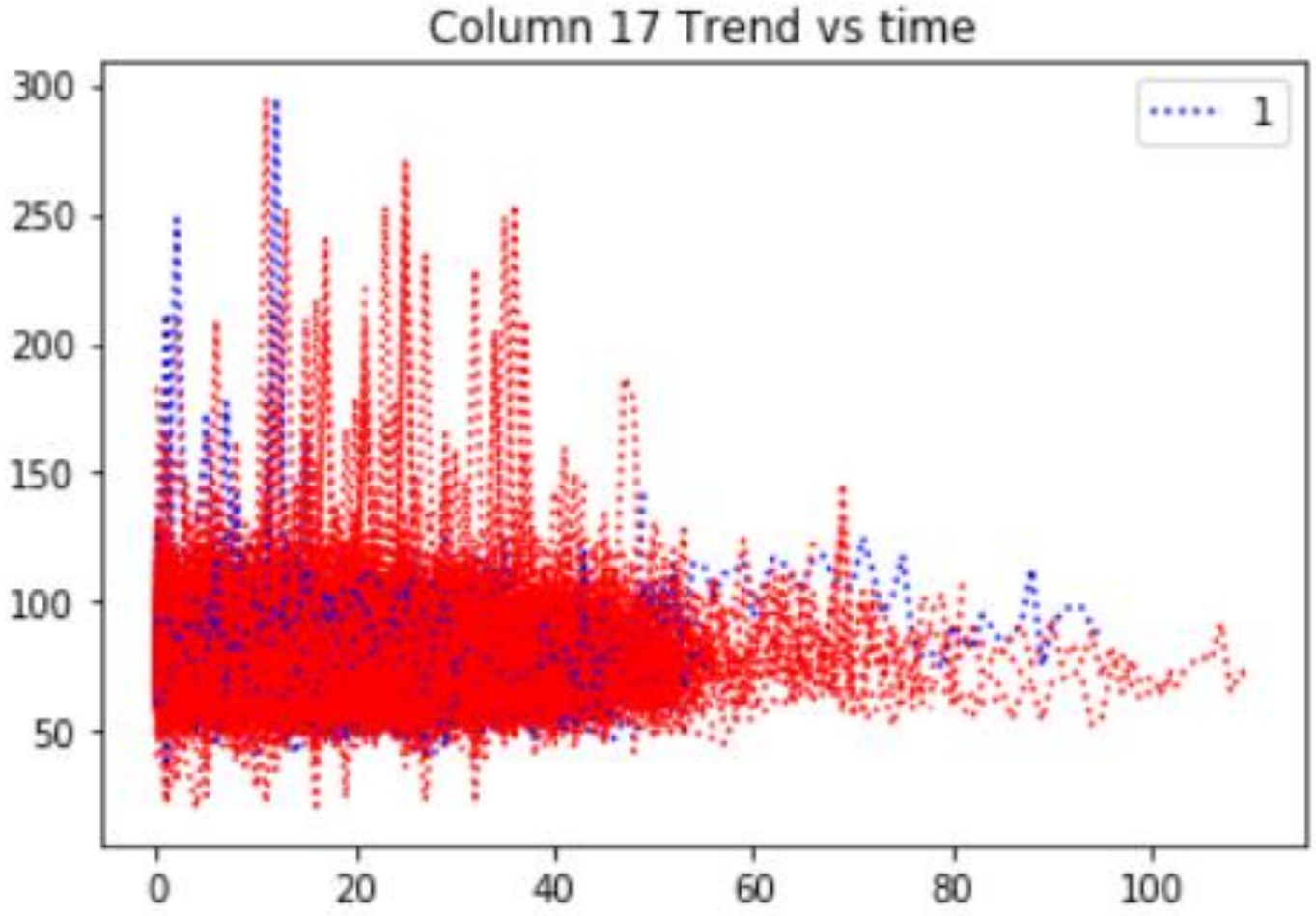



Key takeaway:

- Column 17 is a high variation feature

# DATA EXPLORATION

## TREND VS TIME

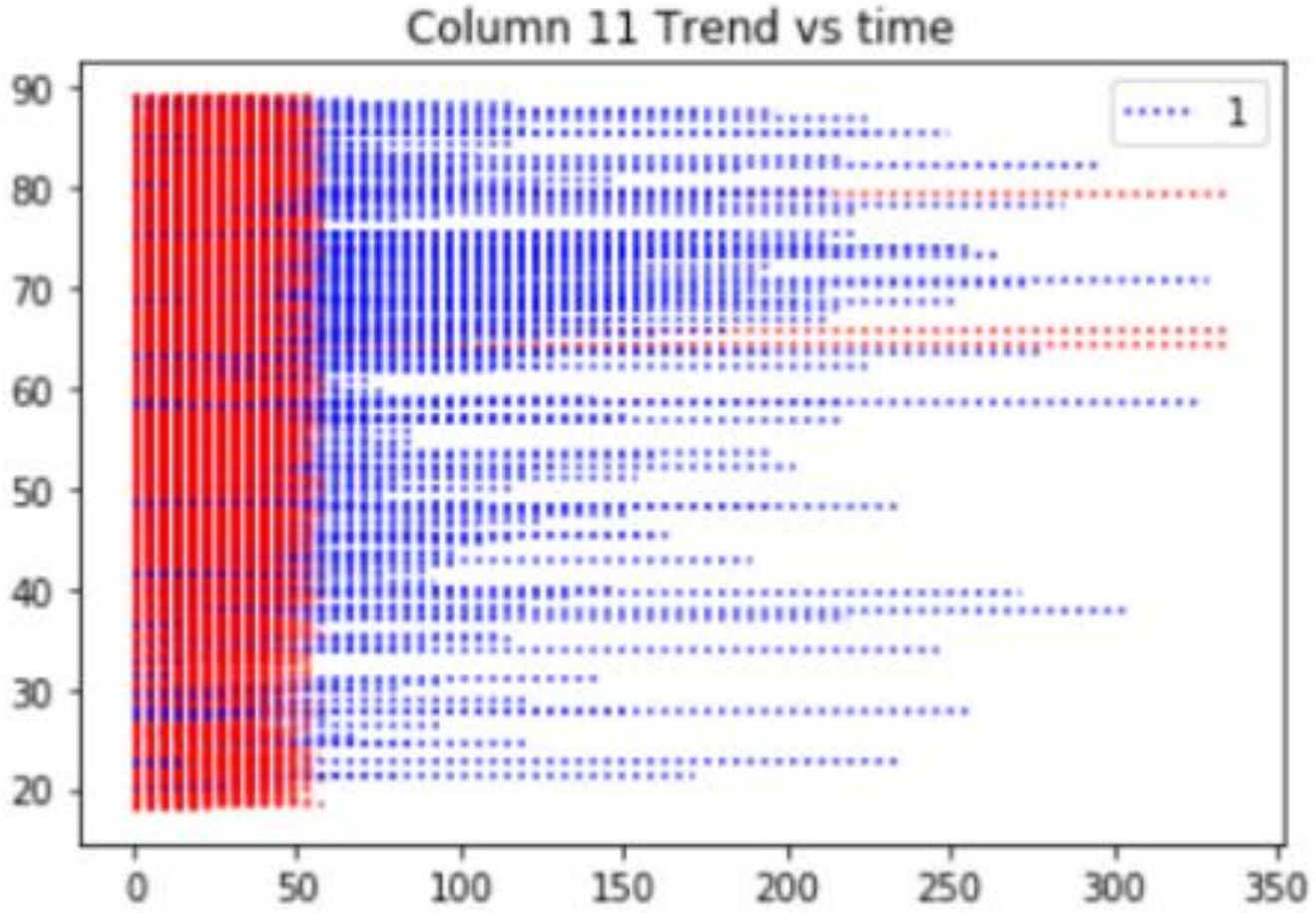



Key takeaway:

- Trend seen from the variance of column 17

# DATA EXPLORATION

## TREND VS TIME




Key takeaway:

- Significant trend from the time length of data for column 11

# DATA PREPROCESSING – NAN HANDLING

- Average value - average value for the column across all the training data.
- 0 – yield worse result for us

| Column | Fill NaN | Preprocessing | Feature Selection | Tuning | Model | Public Score |
|---|---|---|---|---|---|---|
| 5 | 0 | TSFresh | None | None | Gradient Boosting | 0.88856 |
| 5 | Average | TSFresh | None | None | Gradient Boosting | 0.89372 |

# FEATURES EXTRACTION - TSFRESH

- *TSFresh* library is used to extract features of time series data
- Enables us to consider more characteristics of the time series data
- Considers characteristic including standard deviation, variance, mean, count above mean, etc
- Time consuming process, we store it using *pandas to_parquet* function to store the output so that we don't have to re-run this code unless we change how we preprocess the data.

# FEATURES SELECTION – COLUMN SELECTION

| Column | Fill NaN | Preprocessing | Feature Selection | Tuning | Model | Public Score |
|---|---|---|---|---|---|---|
| 18 | Average | TSFresh | None | None | Gradient Boosting | 0.90419 |
| 22 | Average | TSFresh | None | None | Gradient Boosting | 0.91283 |
| 18, 22 | Average | TSFresh | None | None | Gradient Boosting | 0.91187 |
| All | Average | TSFresh | None | None | Gradient Boosting | 0.93715 |
| All but 18 | Average | TSFresh | None | None | Gradient Boosting | 0.94368 |

# FEATURES SELECTION - FEATURE SELECTION TECHNIQUE

| Column | Fill NaN | Preprocessing | Feature Selection | Tuning | Model | Public Score |
|---|---|---|---|---|---|---|
| All but 18 | Average | TSFresh | None | None | Gradient Boosting | 0.94368 |
| All but 18 | Average | TSFresh | Extra Tree Classifier | None | Gradient Boosting | 0.95056 |

# MODEL TRAINING

| Column | Fill NaN | Preprocessing | Feature Selection | Tuning | Model | Public Score |
|---|---|---|---|---|---|---|
| 22 | Average | TSFresh | None | None | Gradient Boosting | 0.91283 |
| 22 | 0 | None | None | None | Tensorflow Neural Network | 0.89307 |
| 22 | 0 | None | None | None | Tensorflow Neural Network | 0.8971 |
| 22 | 0 | None | None | None | Tensorflow Neural Network | 0.84987 |

# MODEL TRAINING

| Column | Fill NaN | Preprocessing | Feature Selection | Tuning | Model | Public Score |
|---|---|---|---|---|---|---|
| All but 18 | Average | TSFresh | Extra Tree Classifier | None | Gradient Boosting | 0.95056 |
| All but 18 | Average | TSFresh | Extra Tree Classifier | None | LightGBM | 0.96069 |
| All but 18 | Average | TSFresh | Extra Tree Classifier | None | XGBoost | 0.95403 |

# MODEL TUNING

| Column | Fill NaN | Preprocessing | Feature Selection | Tuning | Model | Public Score |
|---|---|---|---|---|---|---|
| All but 18 | Average | TSFresh | Extra Tree Classifier | None | LightGBM | 0.96069 |
| All but 18 | Average | TSFresh | Extra Tree Classifier | Hyperopt | LightGBM | 0.96169 |

# LATER WORKS

| Column | Fill NaN | Preprocessing | Feature Selection | Tuning | Model | Public Score |
|---|---|---|---|---|---|---|
| All but 18 | Average | TSFresh | Extra Tree Classifier | None | LightGBM | 0.96069 |
| All but 18 | Average | TSFresh | Extra Tree Classifier | Hyperopt | LightGBM | 0.96169 |
| All but 27 | Average | TSFresh | Extra Tree Classifier | None | LightGBM | 0.95609 |
| All but 27 | Average | TSFresh | Extra Tree Classifier Select For Each Feature | None | LightGBM | 0.96063 |
| All but 27 | Average | TSFresh | Extra Tree Classifier Select For Each Feature | Hyperopt | LightGBM | 0.95909 |

# PRIVATE SCORES

| Column | Fill NaN | Preprocessing | Feature Selection | Tuning | Model | Public Score | Private Scores |
|---|---|---|---|---|---|---|---|
| All but 18 | Average | TSFresh | Extra Tree Classifier | None | LightGBM | 0.96069 | 0.93373 |
| All but 18 | Average | TSFresh | Extra Tree Classifier | Hyperopt | LightGBM | 0.96169 | 0.94081 |
| All but 27 | Average | TSFresh | Extra Tree Classifier Select For Each Feature | Hyperopt | LightGBM | 0.95909 | 0.93639 |

# CONCLUSION

- Standing on the shoulders of giants
  - When there are a lot of useful features, it might be better to use a library like TSFresh to extract the features for you.
- Winners never quit, and quitters never win
  - When there are too many features, just keep trying on finding the best feature groups, you will never know when you will find the best one.
- There's no guarantee of winning, even if you are the best players and the best team.
  - Even if we use hyperopt for tuning, there is no guarantee that the model will be better, we just have to keep trying.