# Credit Card Fraud Project

By Allie & Dave

Davis Data Science Club
23 Winter Project
Credit Card Fraud Team

# Goal

Predict the probability of an online credit card transaction being fraudulent, based on different properties of the transactions.

# How does our group work?

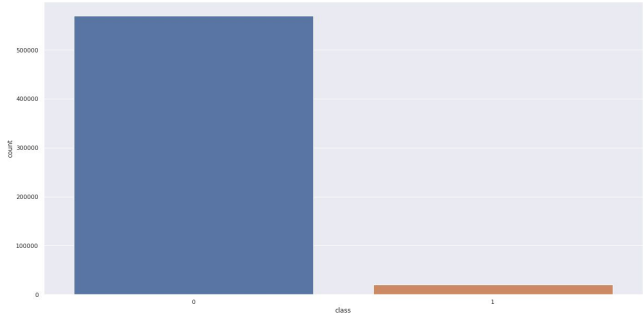
1. Weekly tutorial and office hour
2. Assignments after each tutorial
3. Mainly divided to three groups based on models

# In this presentation

Focus on EDA part

# Data

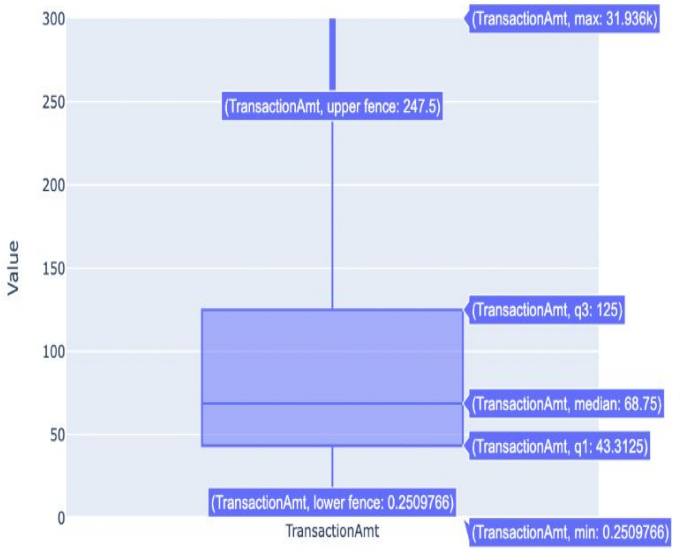
1. **434** features and **590540** observations
2. Only **29** are **categorical variables**, most of them are numerical
3. Some important features: **TransactionDT**, **TransactionAmt**, **Card**, **Product**, **ID**, and etc..
4. Highly **Imbalanced Data**



# Overview of numerical variables

| | TransactionID | isFraud | TransactionDT | TransactionAmt | card1 | card2 | 5 |
|---|---|---|---|---|---|---|---|
| count | 5.905400e+05 | 590540.000000 | 5.905400e+05 | 590540.000000 | 590540.000000 | 581607.000000 | |
| mean | 3.282270e+06 | 0.034990 | 7.372311e+06 | 135.027347 | 9898.734658 | 362.555488 | |
| std | 1.704744e+05 | 0.183755 | 4.617224e+06 | 239.157438 | 4901.170153 | 157.793246 | |
| min | 2.987000e+06 | 0.000000 | 8.640000e+04 | 0.250977 | 1000.000000 | 100.000000 | |
| 25% | 3.134635e+06 | 0.000000 | 3.027058e+06 | 43.312500 | 6019.000000 | 214.000000 | |
| 50% | 3.282270e+06 | 0.000000 | 7.306528e+06 | 68.750000 | 9678.000000 | 361.000000 | |
| 75% | 3.429904e+06 | 0.000000 | 1.124662e+07 | 125.000000 | 14184.000000 | 512.000000 | |
| max | 3.577539e+06 | 1.000000 | 1.581113e+07 | 31936.000000 | 18396.000000 | 600.000000 | |

Boxplot for TransactionAmt



# Feature Engineering & EDA

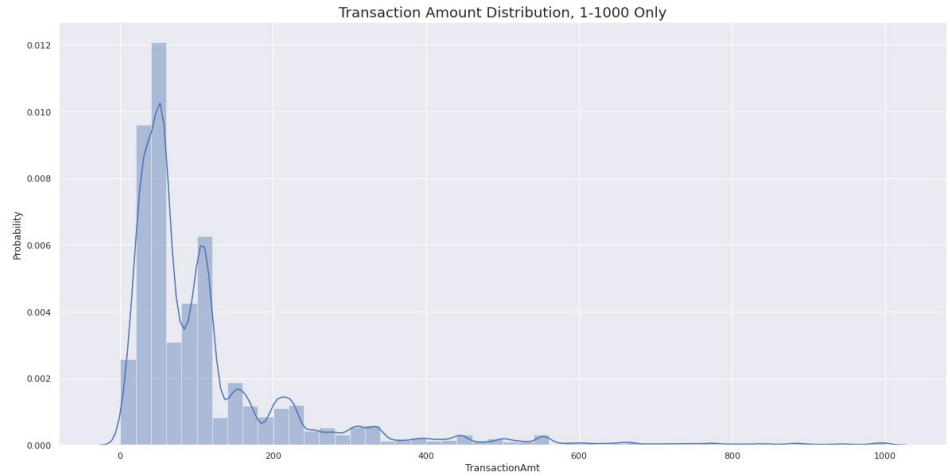
1. Drop columns which has more than 90% null value

```python
# Drop the columns where one category contains more than 90% values
drop_cols = [] # list data structure

# create a for-loop to run through
for col in df.columns:
    missing_share = df[col].isnull().sum()/df.shape[0]
    if missing_share > 0.9:
        drop_cols.append(col)
        print(col)
        # df[col + "_missing_flag"] = df[col].isnull()

good_cols = [col for col in df.columns if col not in drop_cols] # don't want to drop / or keep
```

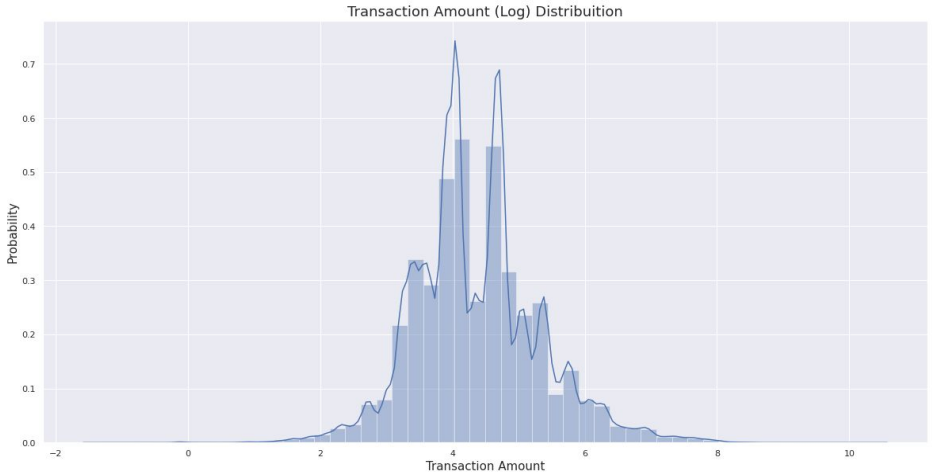
```
dist2
D7
id_07
id_08
id_18
id_21
id_22
id_23
id_24
id_25
id_26
id_27
```

# Feature Engineering & EDA

2. Log transformation on ***TransactionAmt*** variables




Right-skewed

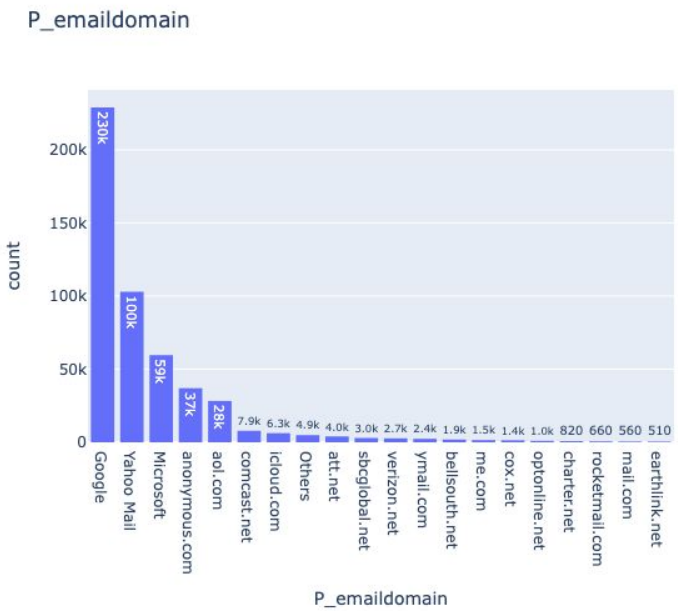



After log-transformation

# Feature Engineering & EDA

3. Change domain's name in ***Email Features***

- **Yahoo Mail:** yahoo.com.mx, yahoo.co.uk, yahoo.fr, yahoo.es
- **Microsoft**: hotmail.com, outlook.com, msn.com, etc…

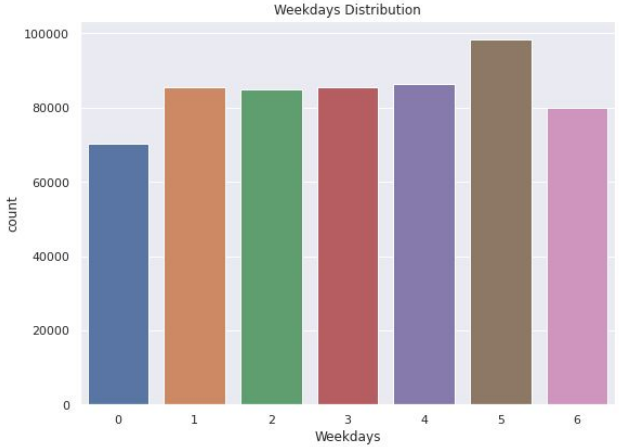


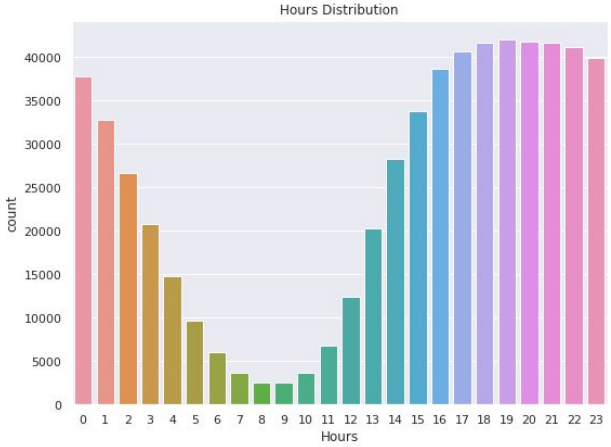

| P_emaildomain | |
|---|---|
| mail.com | 0.189624 |
| Microsoft | 0.053298 |
| Google | 0.043496 |
| icloud.com | 0.031434 |
| comcast.net | 0.031187 |
| charter.net | 0.030637 |
| NoInf | 0.029538 |
| bellsouth.net | 0.027763 |
| Others | 0.025646 |
| anonymous.com | 0.023217 |
| Yahoo Mail | 0.022544 |
| aol.com | 0.021811 |
| earthlink.net | 0.021401 |
| ymail.com | 0.020868 |
| cox.net | 0.020818 |
| me.com | 0.017740 |
| optonline.net | 0.016815 |
| verizon.net | 0.008133 |
| att.net | 0.007439 |
| sbcglobal.net | 0.004040 |
| rocketmail.com | 0.003012 |

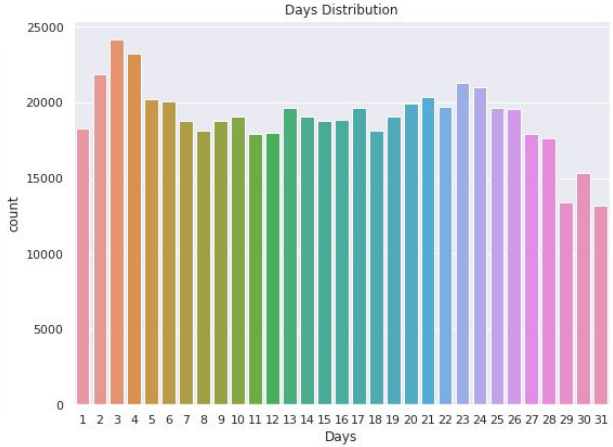
Name: isFraud, dtype: float64

# Feature Engineering & EDA

4. Process ***datetime*** to date, weekdays and hours

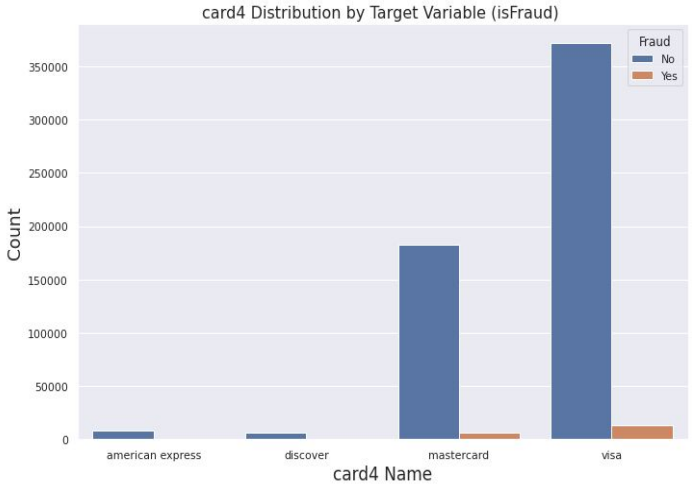


# Feature Engineering & EDA

5. ***Card Features***


```
card4
american express     0.028698
discover             0.077282
mastercard           0.034331
visa                 0.034756
Name: isFraud, dtype: float64
```

Higher chance if you have the "discover" card for fruadulous cases. Mastercard and Visa do take more share of fraudulant cases but their percentages are both lower than "discover" cards.

# Feature Engineering & EDA

6. ***ID features***