

# Building Blocks of Data Analytics



**01.** Data Importance

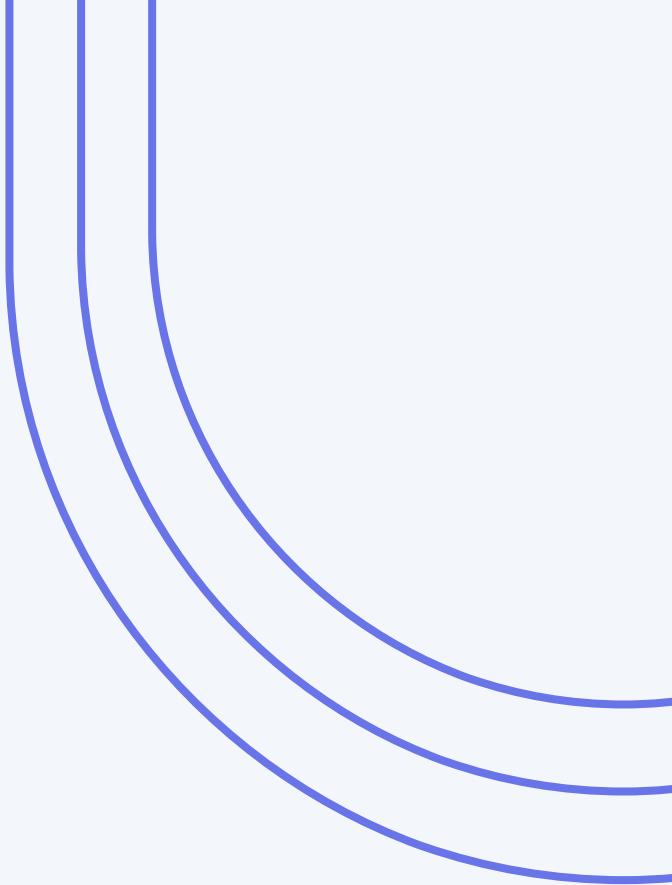
**02.** Data Roles

**03.** Case Study

**04.** Q&A Section

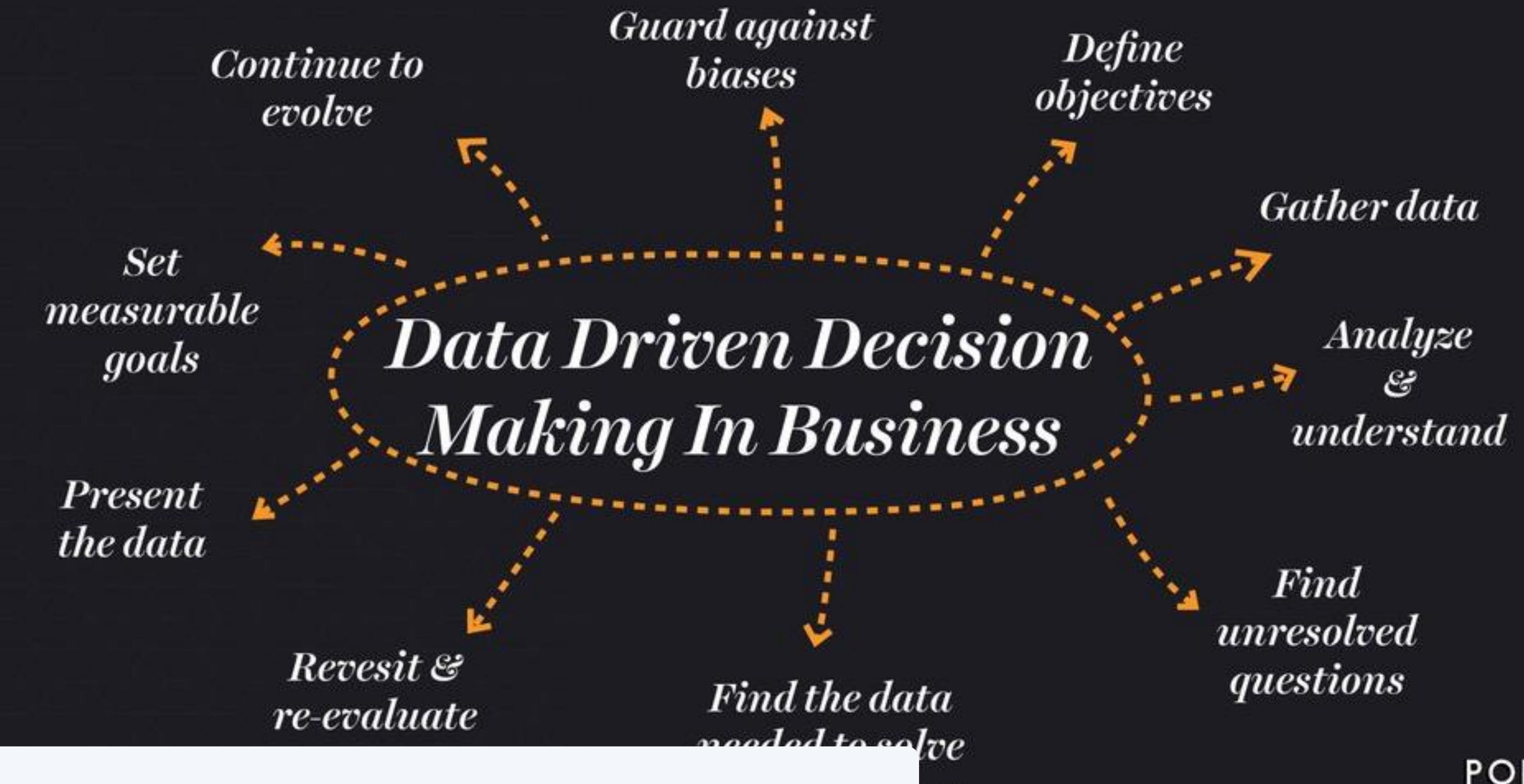


**Agenda**



01.

# Data Importance



Is data really important?



POLICY

# Data Importance

---



## Process Optimization

- It can show company leaders how efficient or costly certain processes are.



## Customer Satisfaction

- Businesses can study the effects of their efforts on customer satisfaction and learn where they can improve. This can help the company create a more pleasing, customized experience for each customer.



## Decision Making

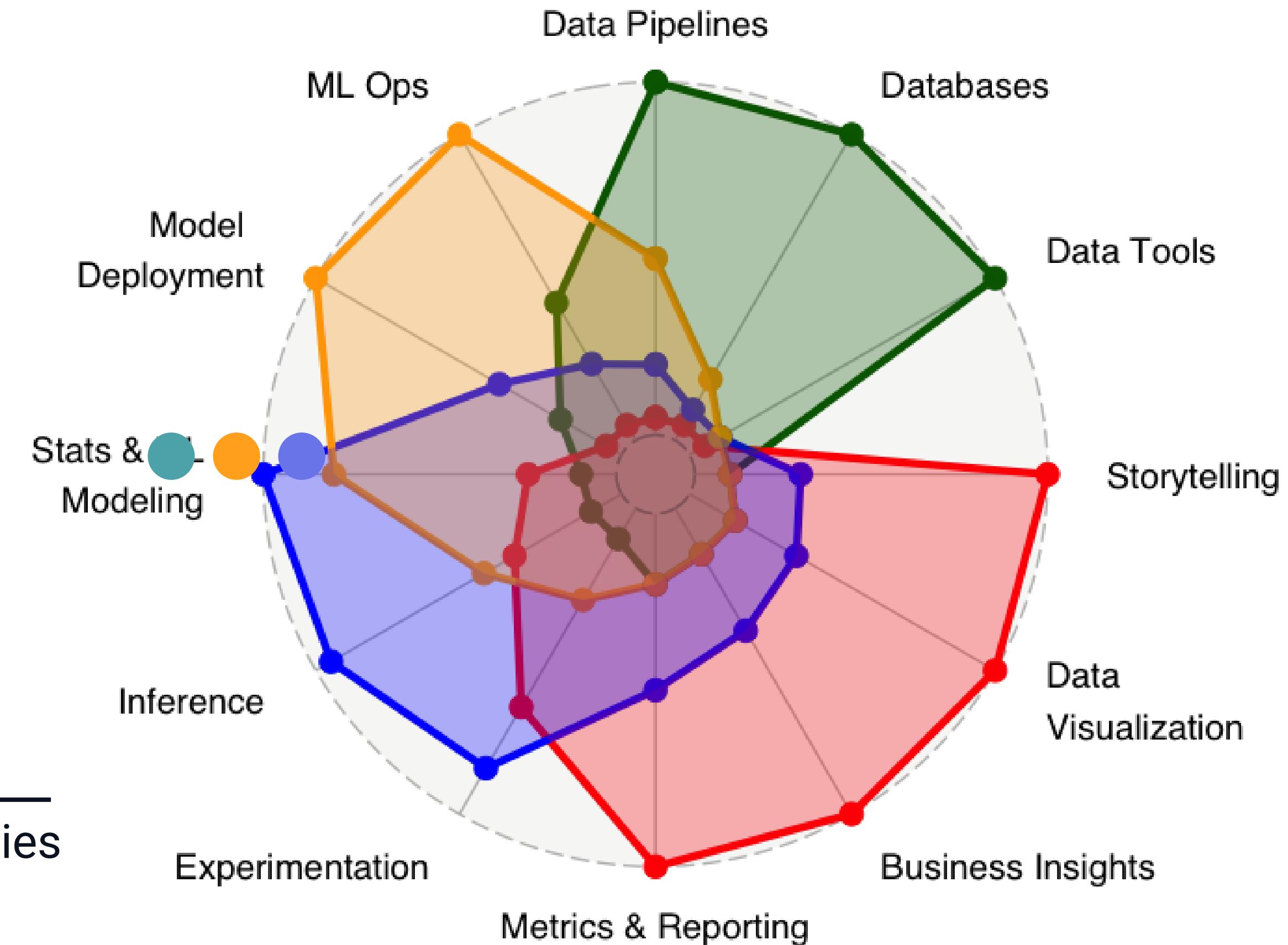
- Help the company make tough decisions more quickly and understand the repercussions or benefits of decisions.

02.

# Data Roles

## Data Roles by Responsibilities

- Data Engineers
- Data Analysts
- Data Scientists
- ML Engineers



03.

# Case Study



Bank



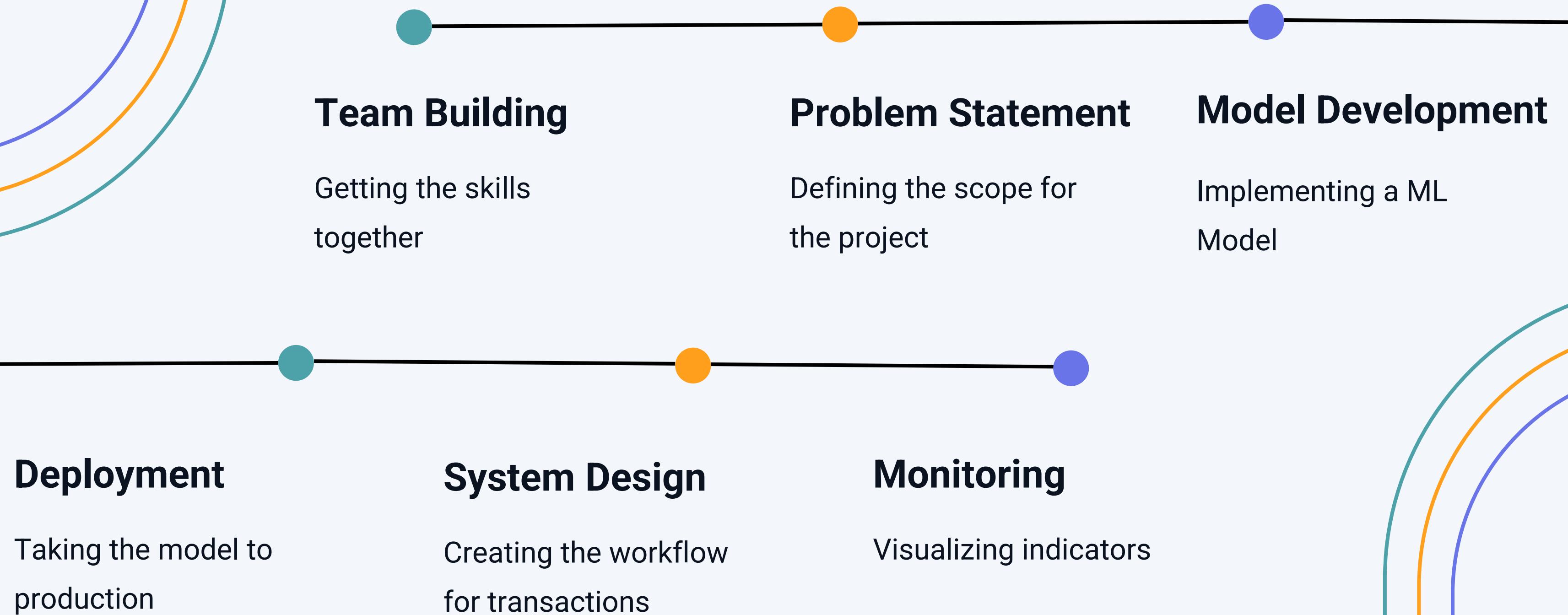
2 out of 5

Credit card transactions  
identified/reported as  
fraudulent

10%

Impact on the bank cost  
regarding associated  
process and irrecoverable

# Project Planning



04.

# The Team

# Dream Team —



The importance of building multidisciplinary teams.



**Logan**

Executive



**James**

Data Scientist



**Olga**

Risk Director



**Brenda**

IT Specialist

05.



# The Problem



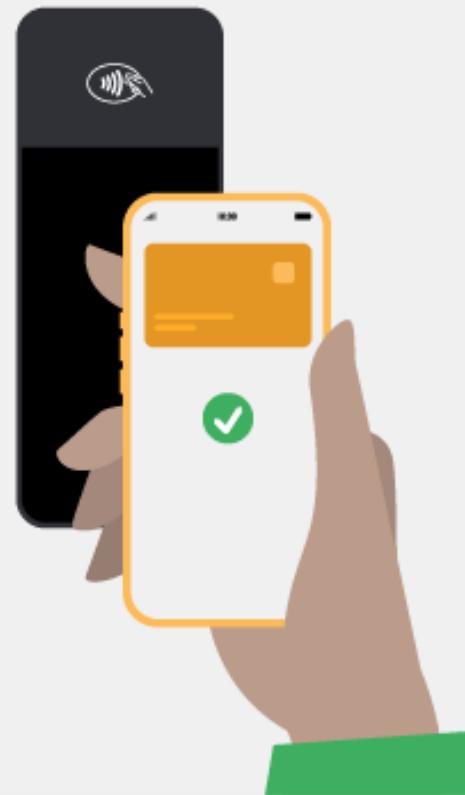
# Fraudulent Transaction

– unauthorized or illegal activity involving the use of payment instruments or financial systems, typically for the purpose of obtaining money, goods, or services without proper consent or authorization from the account holder.

# Transaction Types

## Card Present

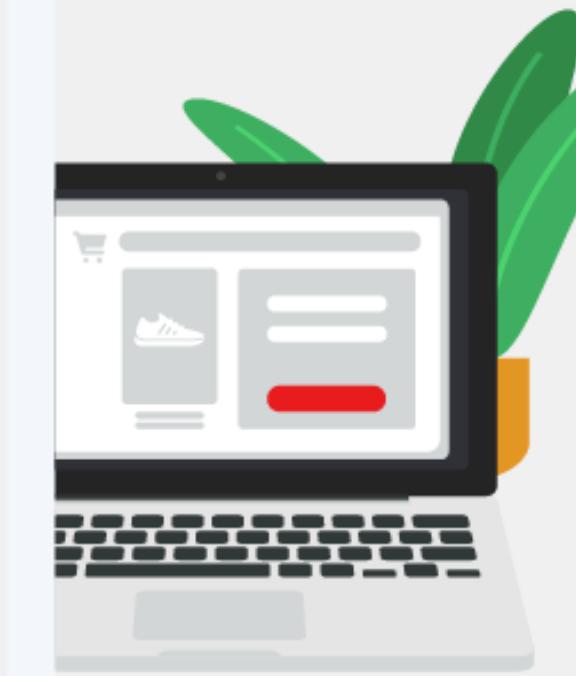
A customer uses their card in a payment terminal.



- Swiping a card
- Inserting a card
- Mobile payments
- Tap-and-go payments

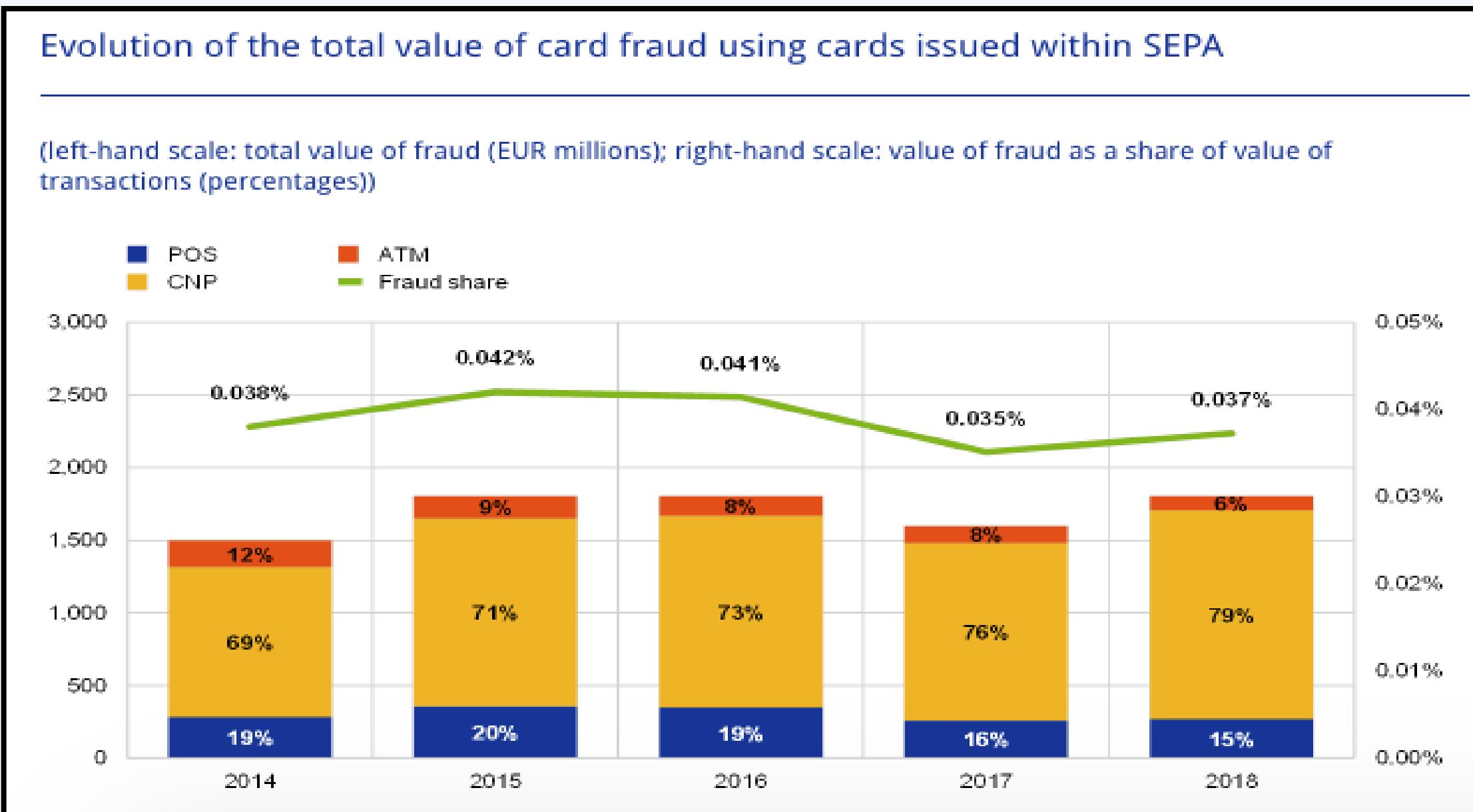
## Card Not Present

Payment is made, but the card is not with the cardholder, or the retailer does not see it.



- Ecommerce
- Mail order and telephone transactions (MOTO)
- Keying a card number in

# Types

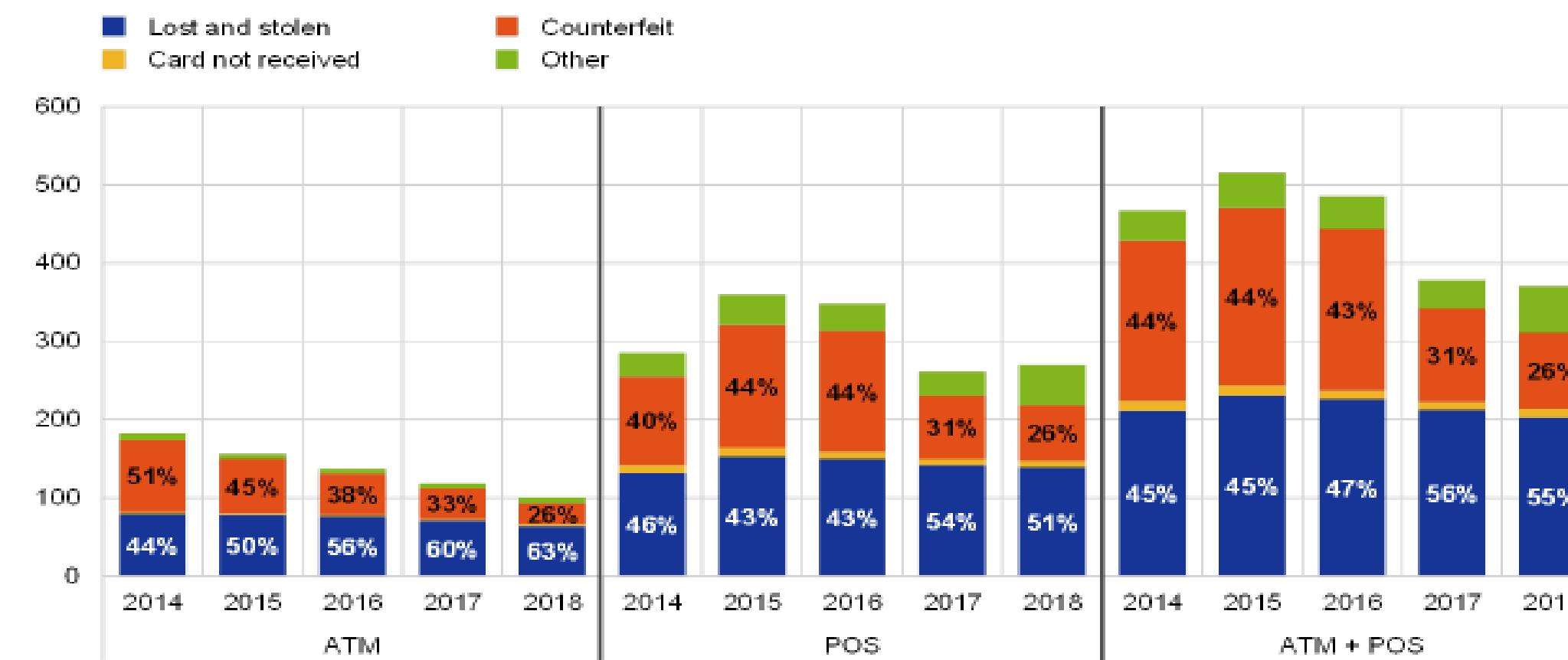


# Types



Evolution and breakdown of the value of card-present fraud by category

(total value of card present fraud (EUR millions))



06.

# The Data

# Credit Card Data

	TX_DATETIME	CUSTOMER_ID	TERMINAL_ID	TX_AMOUNT
0	2018-04-01 07:19:05	0	3	123.59
1	2018-04-01 19:02:02	0	3	46.51
2	2018-04-01 18:00:16	0	0	77.34
3	2018-04-02 15:13:02	0	2	32.35
4	2018-04-02 14:05:38	0	3	63.30
5	2018-04-02 15:46:51	0	3	13.59
6	2018-04-02 08:51:06	0	2	54.72
7	2018-04-02 20:24:47	0	3	51.89
8	2018-04-03 12:15:47	0	2	117.91
9	2018-04-03 08:50:09	0	1	67.72
10	2018-04-03 09:25:49	0	1	28.46
11	2018-04-03 15:33:14	0	2	50.25
12	2018-04-03 07:41:24	0	1	93.26
13	2018-04-04 01:15:35	0	0	46.40
14	2018-04-04 09:33:58	0	2	23.26
15	2018-04-05 16:19:09	0	1	71.96
16	2018-04-05 07:41:19	0	2	52.69

**RowName:** Transaction ID.

**TX\_DateTime:** Transaction date and time.

**Customer\_ID:** Unique identification number of the customer.

**Terminal\_ID:** Unique identification number of the merchant.

**TX\_Amount:** Transaction amount in EUR.



# Credit Card Data



CUSTOMER_ID	x_customer_id	y_customer_id
0	0	54.881350
1	1	42.365480
2	2	96.366276
3	3	56.804456
4	4	2.021840

**Customer\_ID:** Unique identification number of the customer.

**x\_y\_Customer:** Customer's address.

TERMINAL_ID	x_terminal_id	y_terminal_id
0	0	54.881350
1	1	60.276338
2	2	42.365480
3	3	43.758721
4	4	96.366276

**Terminal\_ID:** Unique identification number of the merchant.

**x\_y\_Terminal:** Merchant's address.



07.

# The Model

Selecting the right  
platform for analytics



open source

biztech<sup>age</sup>

**Software  
Licenses**  
5 FAQs About

# Features



	TX_DATETIME	CUSTOMER_ID	TERMINAL_ID	TX_AMOUNT
0	2018-04-01 07:19:05	0	3	123.59
1	2018-04-01 19:02:02	0	3	46.51
2	2018-04-01 18:00:16	0	0	77.34
3	2018-04-02 15:13:02	0	2	32.35
4	2018-04-02 14:05:38	0	3	63.30
5	2018-04-02 15:46:51	0	3	13.59
6	2018-04-02 08:51:06	0	2	54.72
7	2018-04-02 20:24:47	0	3	51.89
8	2018-04-03 12:15:47	0	2	117.91
9	2018-04-03 08:50:09	0	1	67.72
10	2018-04-03 09:25:49	0	1	28.46
11	2018-04-03 15:33:14	0	2	50.25
12	2018-04-03 07:41:24	0	1	93.26
13	2018-04-04 01:15:35	0	0	46.40
14	2018-04-04 09:33:58	0	2	23.26
15	2018-04-05 16:19:09	0	1	71.96
16	2018-04-05 07:41:19	0	2	52.69

Date Centric
Variables
<ul style="list-style-type: none"><li>• Day of Week.</li><li>• Hour of Day.</li></ul>
<ul style="list-style-type: none"><li>• Weekend Flag.</li><li>• Night Flag.</li></ul>

Customer Centric
Variables
<ul style="list-style-type: none"><li>• Average Daily Amount.</li><li>• Total Daily Amount.</li></ul>
<ul style="list-style-type: none"><li>• Average Daily Transactions.</li><li>• Total Daily Transactions.</li></ul>
<ul style="list-style-type: none"><li>• Fraud History.</li><li>• First Time Purchase.</li></ul>

Merchant Centric
Variables
<ul style="list-style-type: none"><li>• Average Daily Amount.</li><li>• Total Daily Amount.</li></ul>
<ul style="list-style-type: none"><li>• Average Daily Transactions.</li><li>• Total Daily Transactions.</li></ul>
<ul style="list-style-type: none"><li>• Fraud History.</li><li>• Clientele</li></ul>

# Explore



```
tibble [1,754,155 x 20] (s3: tbl_df/tbl/data.frame)
$ TRANS_FRAUD      : num [1:1754155] 0 0 0 0 0 0 0 0 0 ...
$ CUSTOMER_ID      : num [1:1754155] 0 0 0 0 0 0 0 0 0 ...
$ TERMINAL_ID      : num [1:1754155] 29 29 29 29 29 87 87 87 87 ...
$ TRANS_AMOUNT     : num [1:1754155] 83.2 95.9 28.5 50.8 39 ...
$ TRANS_DAY_OF_WEEK: num [1:1754155] 6 3 3 2 3 5 3 2 1 3 ...
$ TRANS_HOUR_OF_DAY: int [1:1754155] 22 13 11 13 11 12 7 5 19 12 ...
$ TRANS_WEEKEND    : num [1:1754155] 1 0 0 0 0 0 0 0 0 ...
$ TRANS_NIGHT      : num [1:1754155] 1 0 0 0 0 0 0 1 0 0 ...
$ CT_AVG_DAY_AMOUNT: num [1:1754155] 147 147 147 147 147 ...
$ CT_AVG_DAY_TRANS : num [1:1754155] 2.4 2.4 2.4 2.4 2.4 2.4 2.4 2.4 2.4 ...
$ CT_TOT_DAY_AMOUNT: num [1:1754155] 132 237 128 251 148 ...
$ CT_TOT_DAY_TRANS : int [1:1754155] 2 4 2 3 2 3 2 1 3 2 ...
$ CT_FRAUD_HIST   : num [1:1754155] 0 0 0 0 0 0 0 0 0 ...
$ TM_AVG_DAY_AMOUNT: num [1:1754155] 105 105 105 105 105 ...
$ TM_AVG_DAY_TRANS : num [1:1754155] 1.58 1.58 1.58 1.58 1.58 ...
$ TM_TOT_DAY_AMOUNT: num [1:1754155] 83.2 119.1 295.7 50.8 39 ...
$ TM_TOT_DAY_TRANS : int [1:1754155] 1 2 4 1 1 2 4 2 4 1 ...
$ TM_FRAUD_HIST   : num [1:1754155] 0 2 2 2 2 0 0 0 0 ...
$ TM_CLIENTELE    : int [1:1754155] 38 38 38 38 38 34 34 34 34 ...
$ CT_PURCHASE_TM  : num [1:1754155] 0 1 2 3 4 0 1 2 3 4 ...
```

**1,754,155 transactions**  
**20 variables**

A tibble: 6 × 20																			
TRANS_FRAUD	CUSTOMER_ID	TERMINAL_ID	TRANS_AMOUNT	TRANS_DAY_OF_WEEK	TRANS_HOUR_OF_DAY	TRANS_WEEKEND	TRANS_NIGHT	CT_AVG_DAY_AMOUNT	CT_AVG_DAY_TRANS	CT_TOT_DAY_AMOUNT	CT_TOT_DAY_TRANS	CT_FRAUD_HIST	TM_AVG_DAY_AMOUNT	TM_AVG_DAY_TRANS	TM_TOT_DAY_AMOUNT	TM_TOT_DAY_TRANS	TM_FRAUD_HIST	TM_CLIENTELE	CT_PURCHASE_TM
<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<int>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
0	0	29	83.19	6	22	1	1	147.477	2.4	147.477	2.4	0	147.477	2.4	147.477	2.4	0	38	0
0	0	29	95.94	3	13	0	0	147.477	2.4	147.477	2.4	0	147.477	2.4	147.477	2.4	0	38	0
0	0	29	28.46	3	11	0	0	147.477	2.4	147.477	2.4	0	147.477	2.4	147.477	2.4	0	38	0
0	0	29	50.75	2	13	0	0	147.477	2.4	147.477	2.4	0	147.477	2.4	147.477	2.4	0	38	0
0	0	29	39.04	3	11	0	0	147.477	2.4	147.477	2.4	0	147.477	2.4	147.477	2.4	0	38	0
0	0	87	33.14	5	12	0	0	147.477	2.4	147.477	2.4	0	147.477	2.4	147.477	2.4	0	38	0

5 rows | 1-10 of 20 columns

# Explore

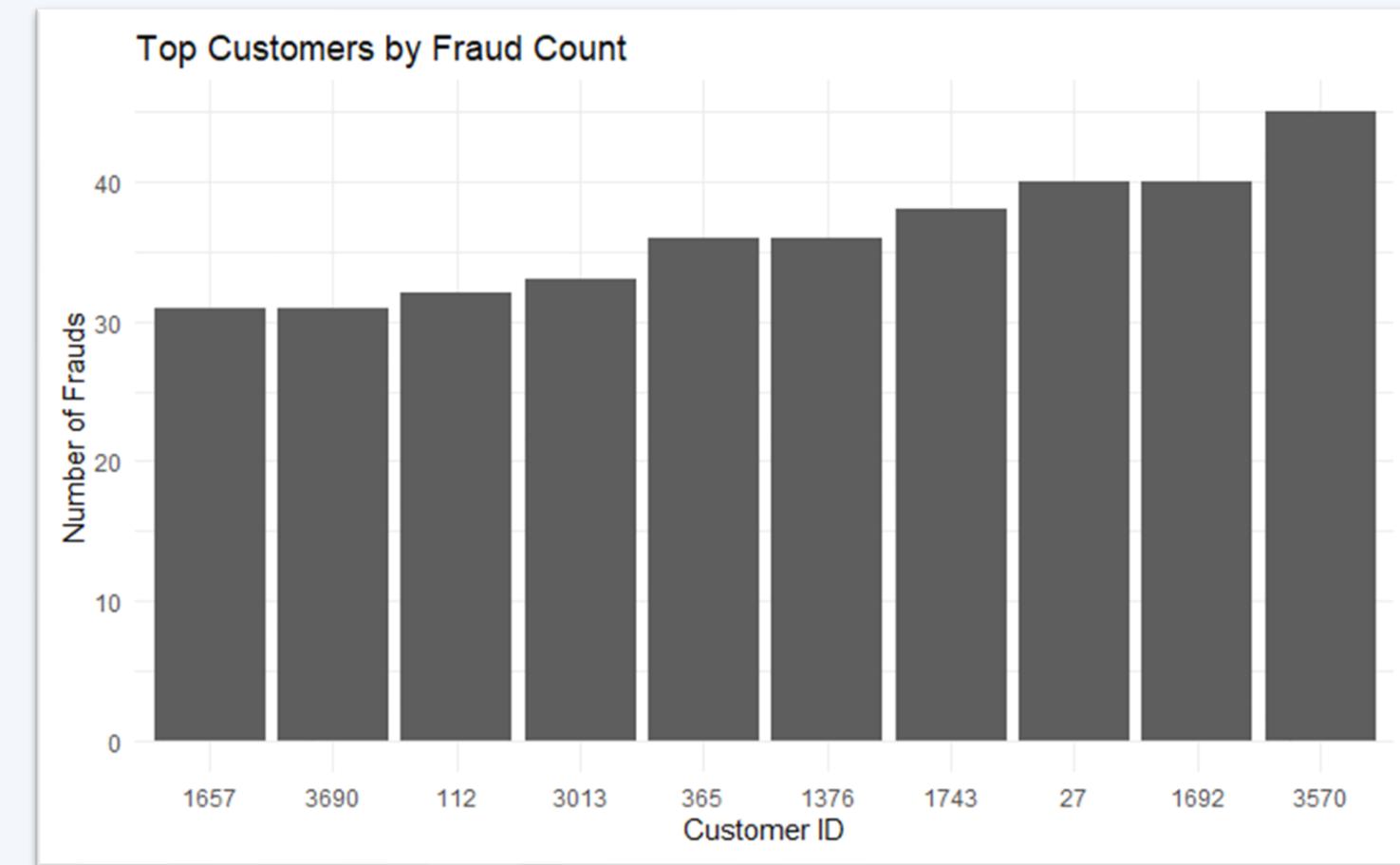
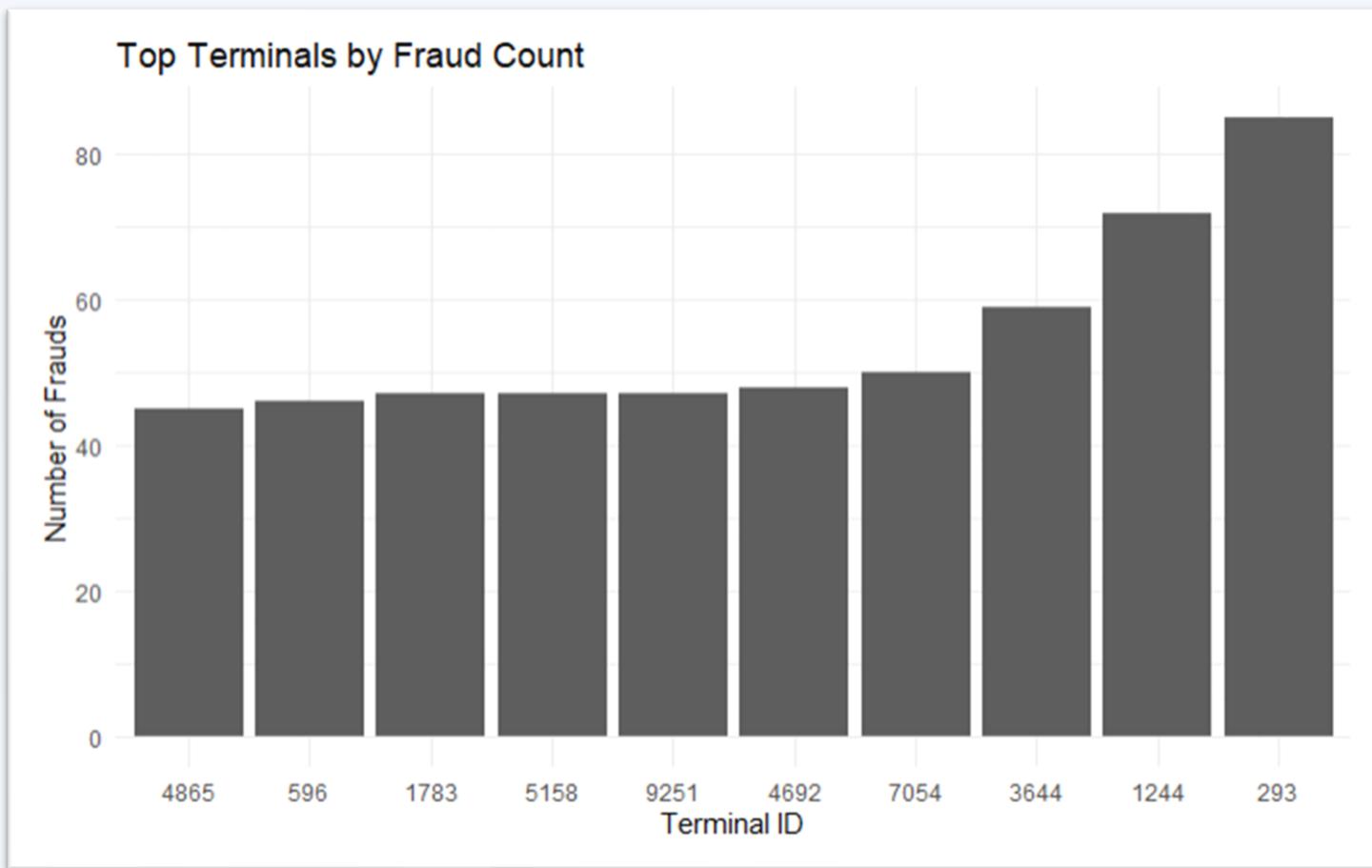


TRANS_ID	DATETIME	CUSTOMER_ID	TERMINAL_ID	TRANS_AMOUNT	TRANS_FRAUD	TRANS_FRAUD_SCENARIO	TRANS_DAY_OF_WEEK	TRANS_HOUR_OF_DAY	TRANS_WEEKEND
Min. : 0	Min. :2018-04-01 00:00:31.00	Min. : 0	Min. : 0	Min. : 0.00	Min. :0.000000	Min. :0.00000	Min. :1.000	Min. : 0.0	Min. :0.0000
1st Qu.: 438539	1st Qu.:2018-05-16 14:40:46.50	1st Qu.:1252	1st Qu.:2502	1st Qu.: 21.01	1st Qu.:0.000000	1st Qu.:0.00000	1st Qu.:2.000	1st Qu.: 8.0	1st Qu.:0.0000
Median : 877077	Median :2018-07-01 11:11:10.00	Median :2506	Median :4994	Median : 44.64	Median :0.000000	Median :0.00000	Median :4.000	Median :11.0	Median :0.0000
Mean : 877077	Mean :2018-07-01 11:20:33.71	Mean :2504	Mean :4997	Mean : 53.63	Mean :0.008369	Mean :0.01882	Mean :3.981	Mean :11.5	Mean :0.2837
3rd Qu.:1315616	3rd Qu.:2018-08-16 08:01:01.50	3rd Qu.:3765	3rd Qu.:7495	3rd Qu.: 76.95	3rd Qu.:0.000000	3rd Qu.:0.00000	3rd Qu.:6.000	3rd Qu.:15.0	3rd Qu.:1.0000
Max. :1754154	Max. :2018-09-30 23:59:57.00	Max. :4999	Max. :9999	Max. :2628.00	Max. :1.000000	Max. :3.00000	Max. :7.000	Max. :23.0	Max. :1.0000
TRANS_NIGHT	LOG_TRANS_AMOUNT	TRANS_AMOUNT_BIN	CT_AVG_DAY_AMOUNT	CT_AVG_DAY_TRANS	CT_TOT_DAY_AMOUNT	CT_TOT_DAY_TRANS	CT_FRAUD_HIST	TM_AVG_DAY_AMOUNT	TM_AVG_DAY_TRANS
Min. :0.0000	Min. :0.000	low :438733	Min. : 5.383	Min. :1.000	Min. : 0.00	Min. : 1.00	Min. : 0.000	Min. : 44.27	Min. :1.068
1st Qu.:0.0000	1st Qu.:3.091	medium :438492	1st Qu.: 73.708	1st Qu.:2.235	1st Qu.: 67.88	1st Qu.: 2.00	1st Qu.: 0.000	1st Qu.: 76.19	1st Qu.:1.478
Median :0.0000	Median :3.821	high :438467	Median :140.048	Median :2.931	Median : 147.24	Median : 3.00	Median : 0.000	Median : 84.71	Median :1.574
Mean :0.1897	Mean :3.660	very_high:438463	Mean :151.685	Mean :2.827	Mean : 191.61	Mean : 3.57	Mean : 1.857	Mean : 85.20	Mean :1.589
3rd Qu.:0.0000	3rd Qu.:4.356		3rd Qu.:217.138	3rd Qu.:3.469	3rd Qu.: 269.91	3rd Qu.: 5.00	3rd Qu.: 2.000	3rd Qu.: 93.56	3rd Qu.:1.686
Max. :1.0000	Max. :7.874		Max. :433.615	Max. :4.281	Max. :4254.52	Max. :14.00	Max. :45.000	Max. :141.32	Max. :2.426
TM_TOT_DAY_AMOUNT	TM_TOT_DAY_TRANS	TM_FRAUD_HIST	TM_CLIENTELE	CT_PURCHASE_TM					
Min. : 0.00	Min. : 1.000	Min. : 0.0000	Min. :10.00	Min. : 0.000					
1st Qu.: 46.28	1st Qu.: 1.000	1st Qu.: 0.0000	1st Qu.:31.00	1st Qu.: 1.000					
Median : 90.49	Median : 2.000	Median : 0.0000	Median :35.00	Median : 2.000					
Mean : 107.26	Mean : 2.001	Mean : 0.7327	Mean :34.55	Mean : 3.226					
3rd Qu.: 148.93	3rd Qu.: 3.000	3rd Qu.: 0.0000	3rd Qu.:39.00	3rd Qu.: 5.000					
Max. :2628.00	Max. :11.000	Max. :84.0000	Max. :59.00	Max. :35.000					

## Summary Statistics

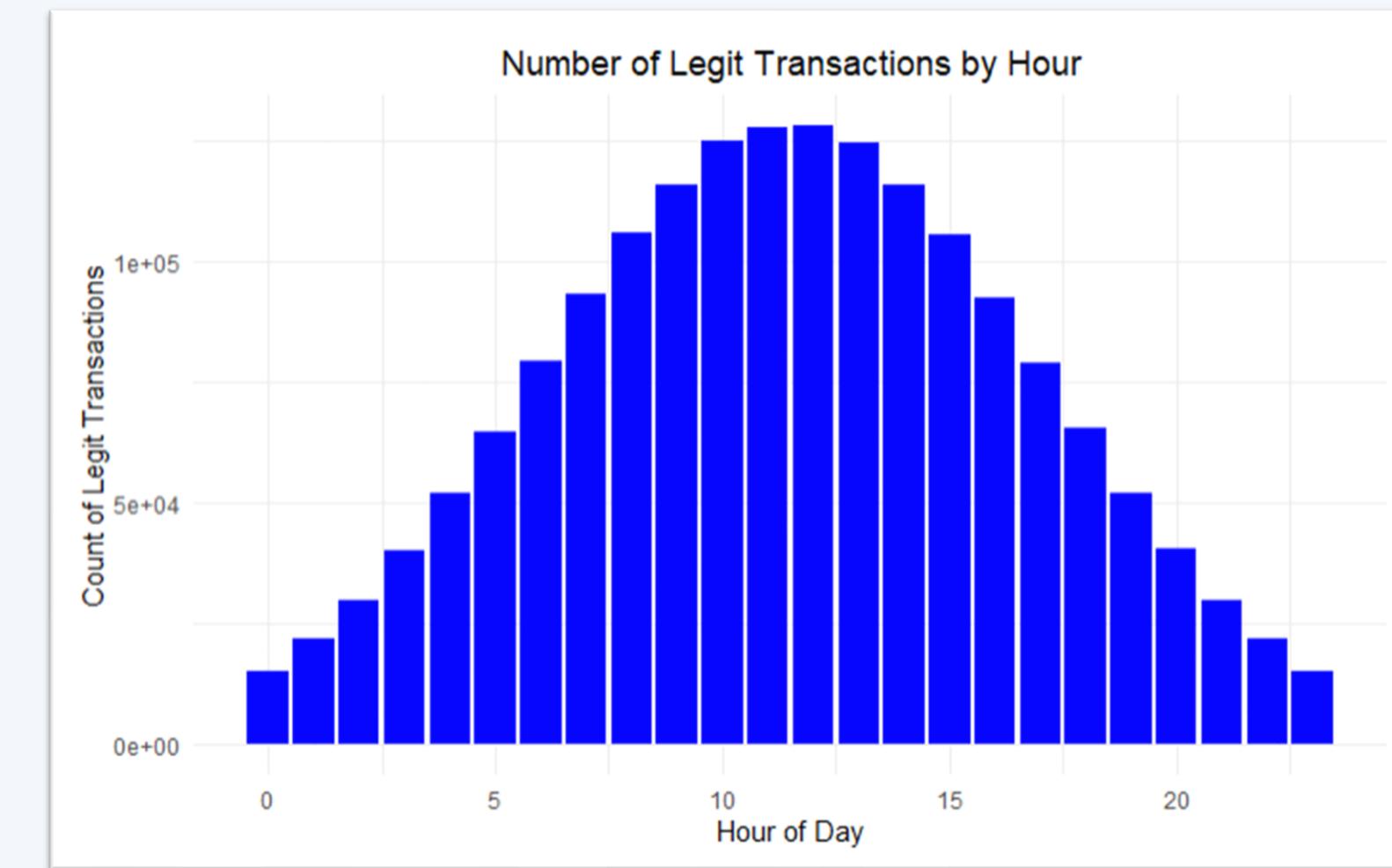
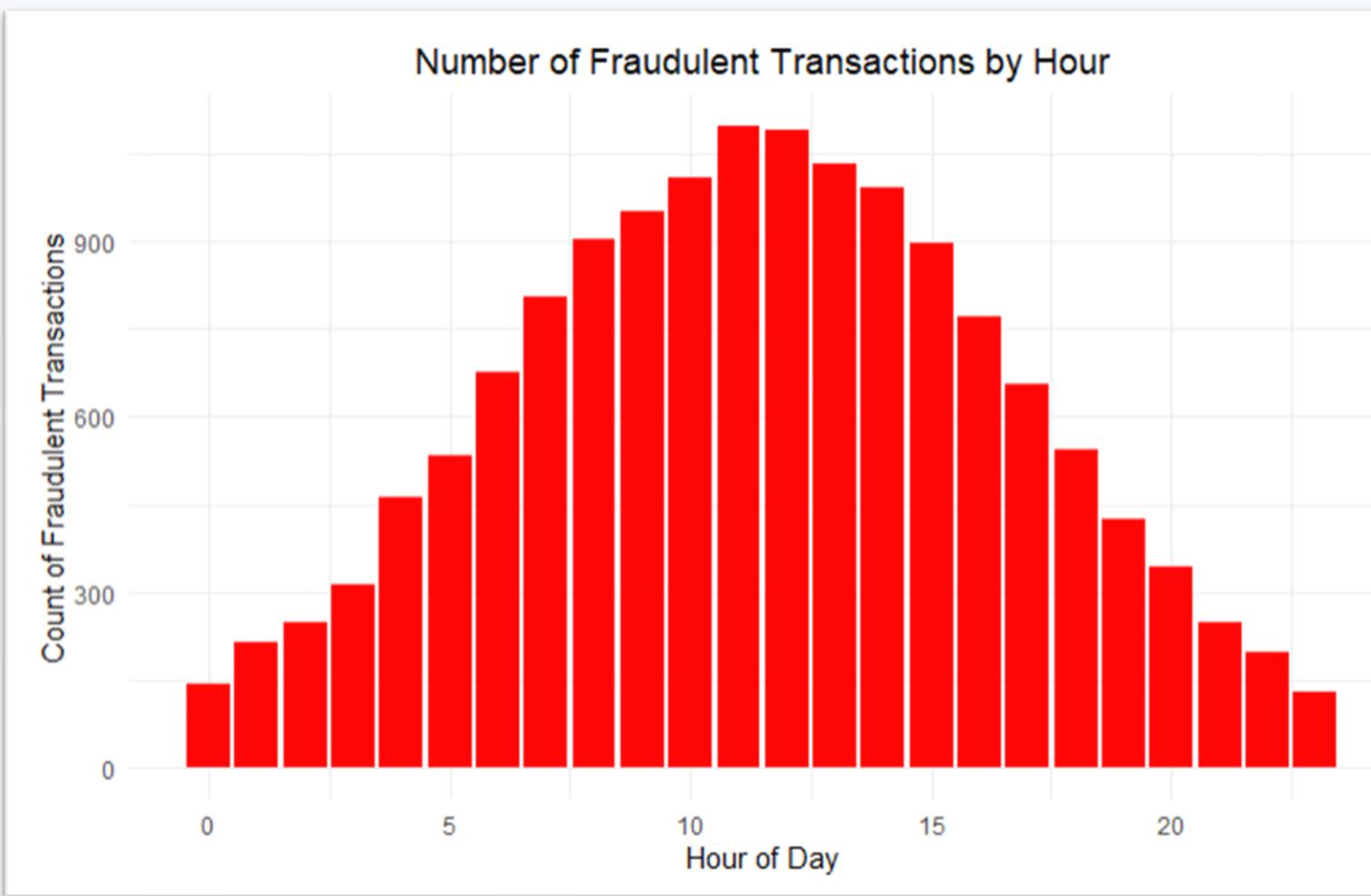
# Explore

---



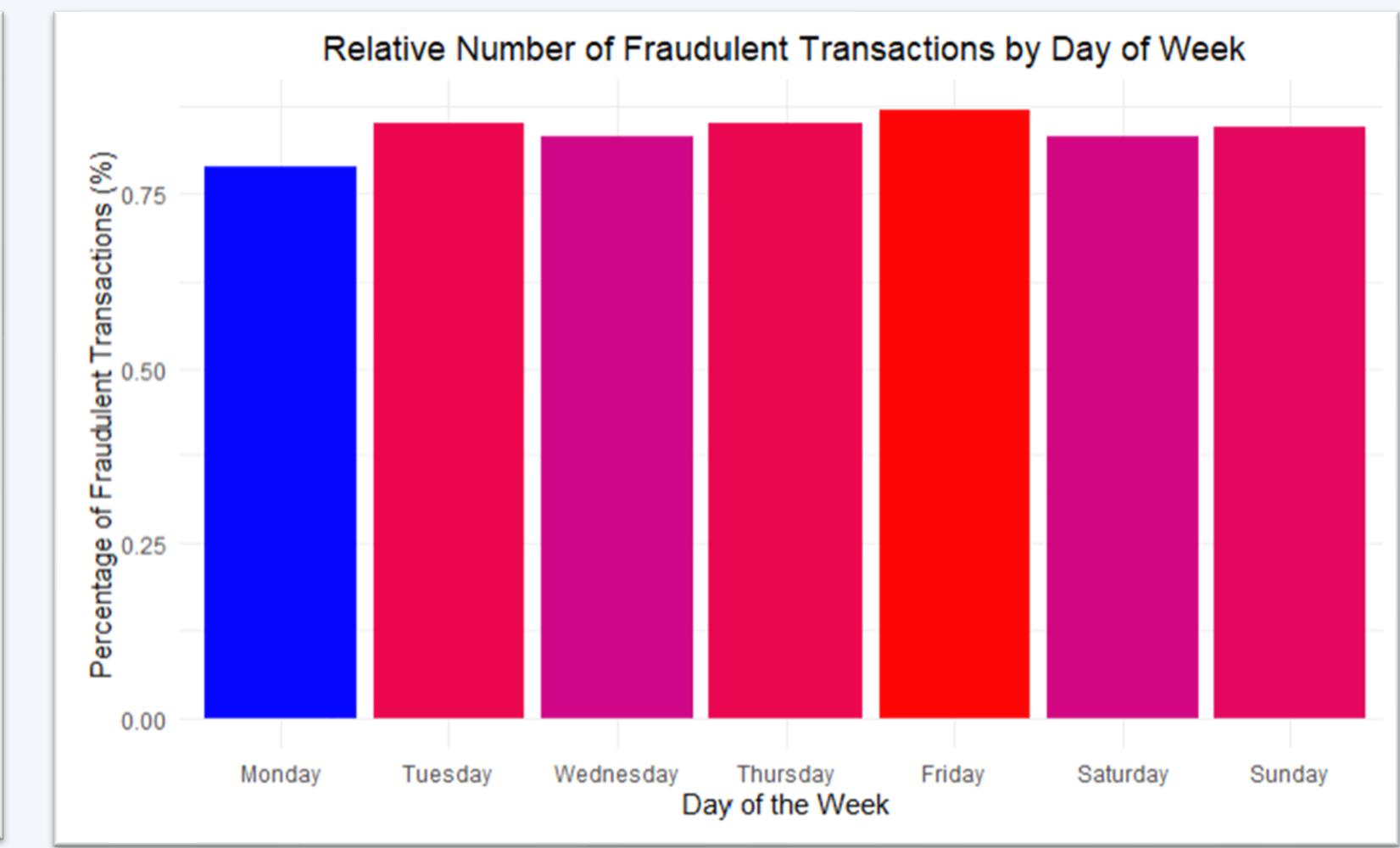
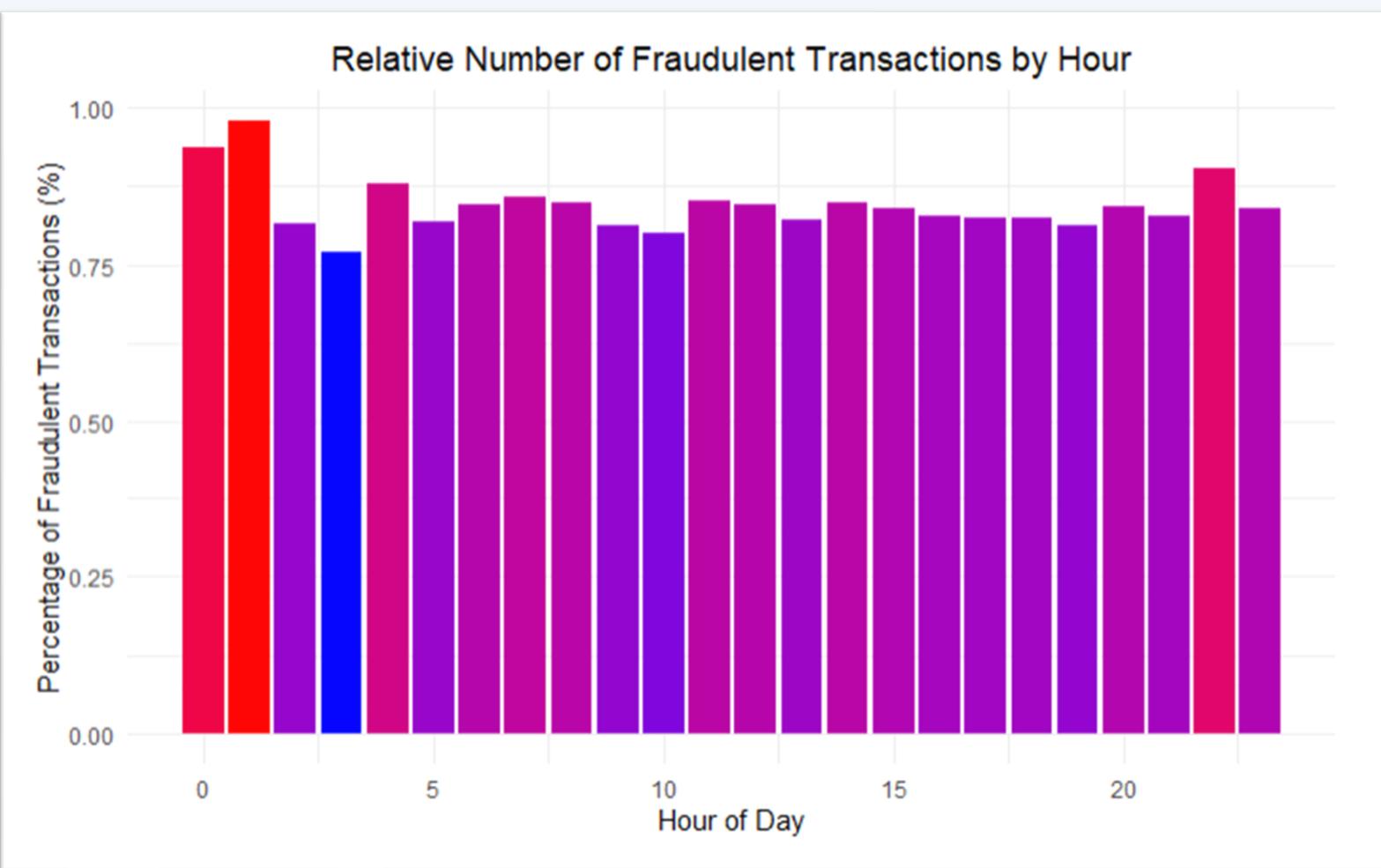
# Explore

---



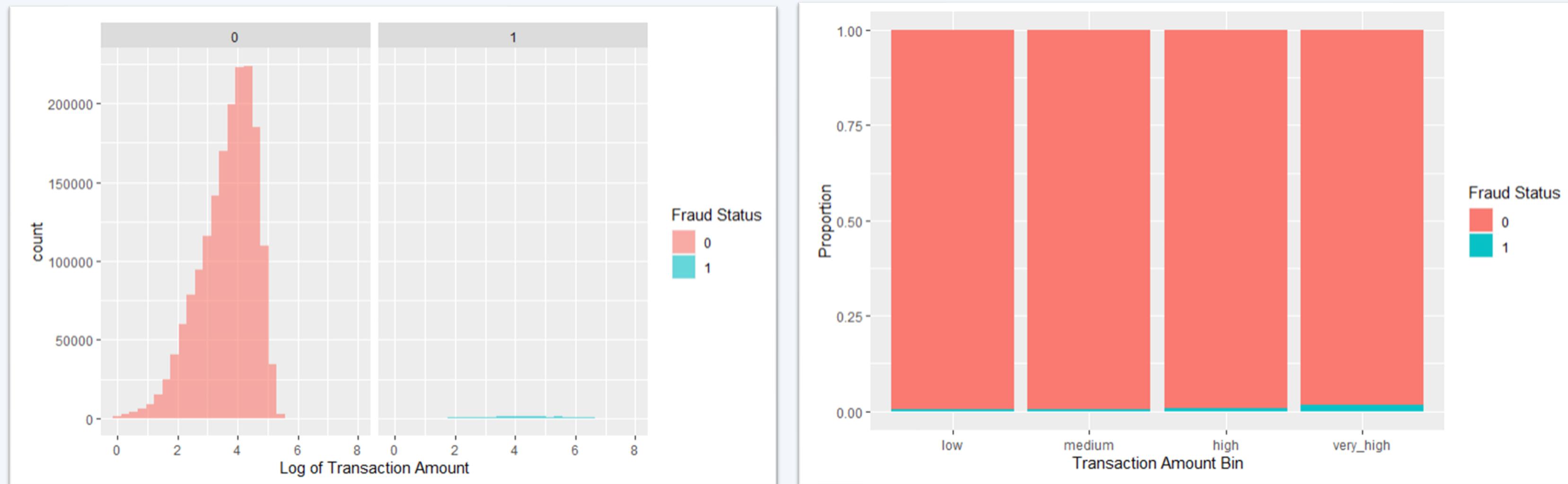
# Explore

---

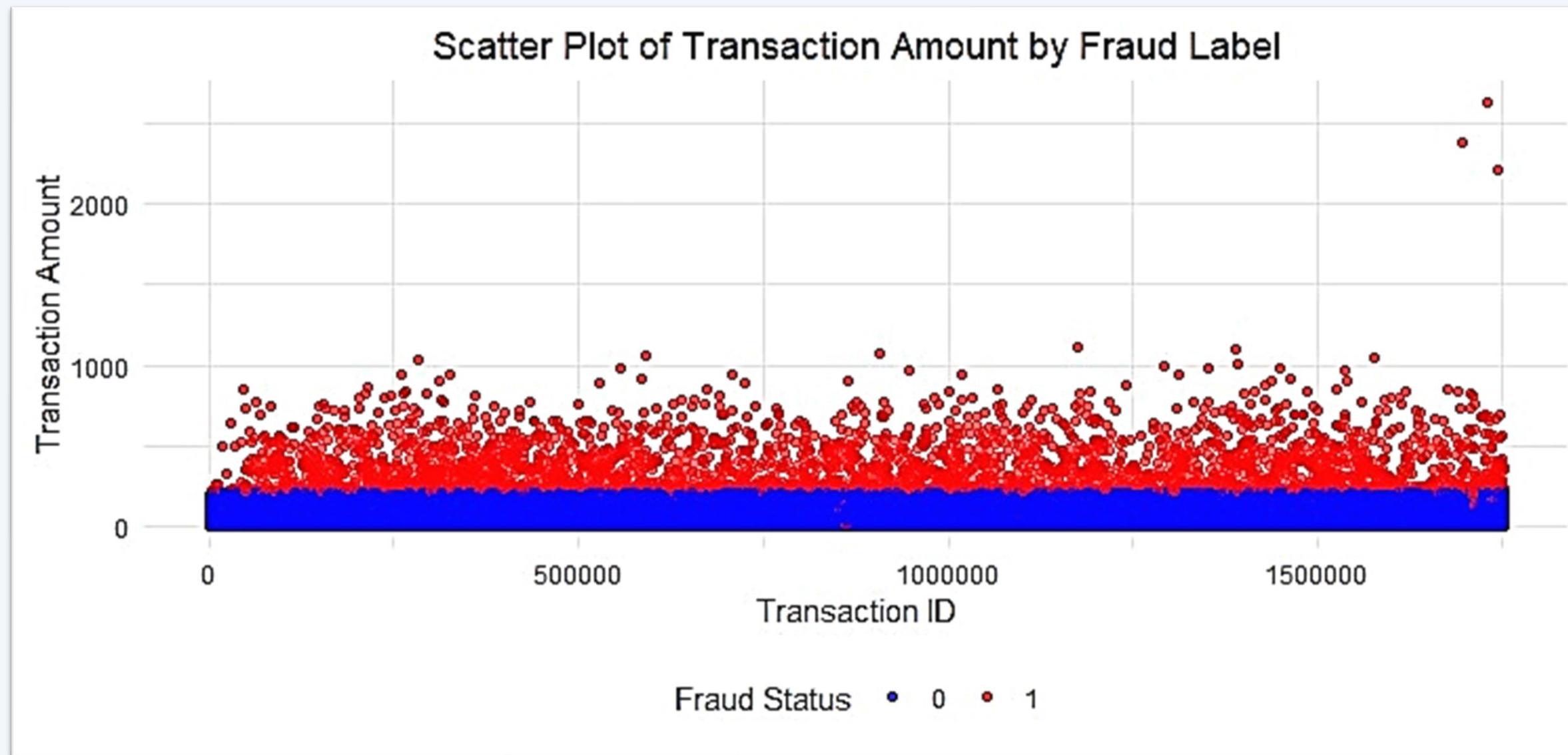


# Explore

---

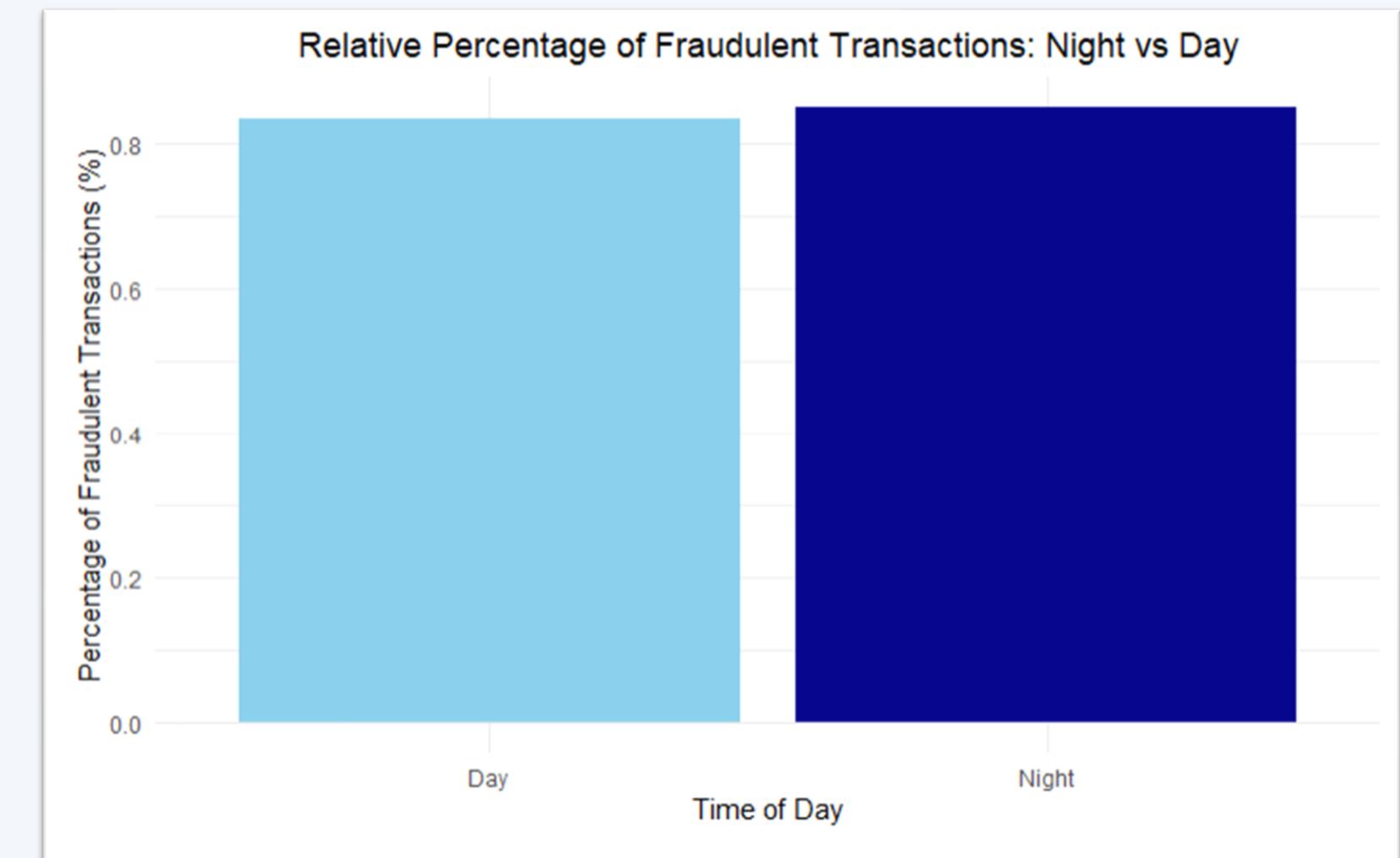
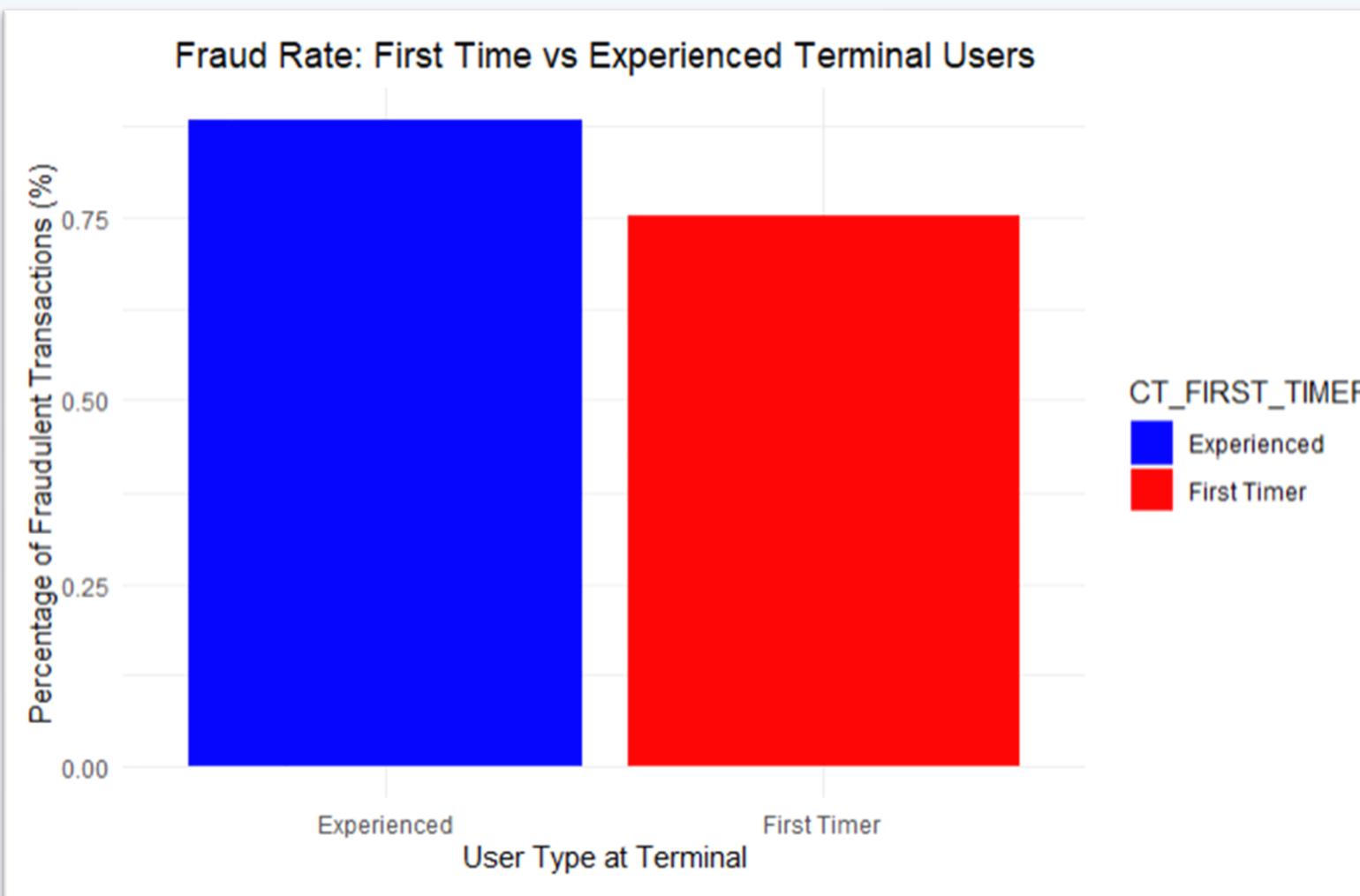


# Explore



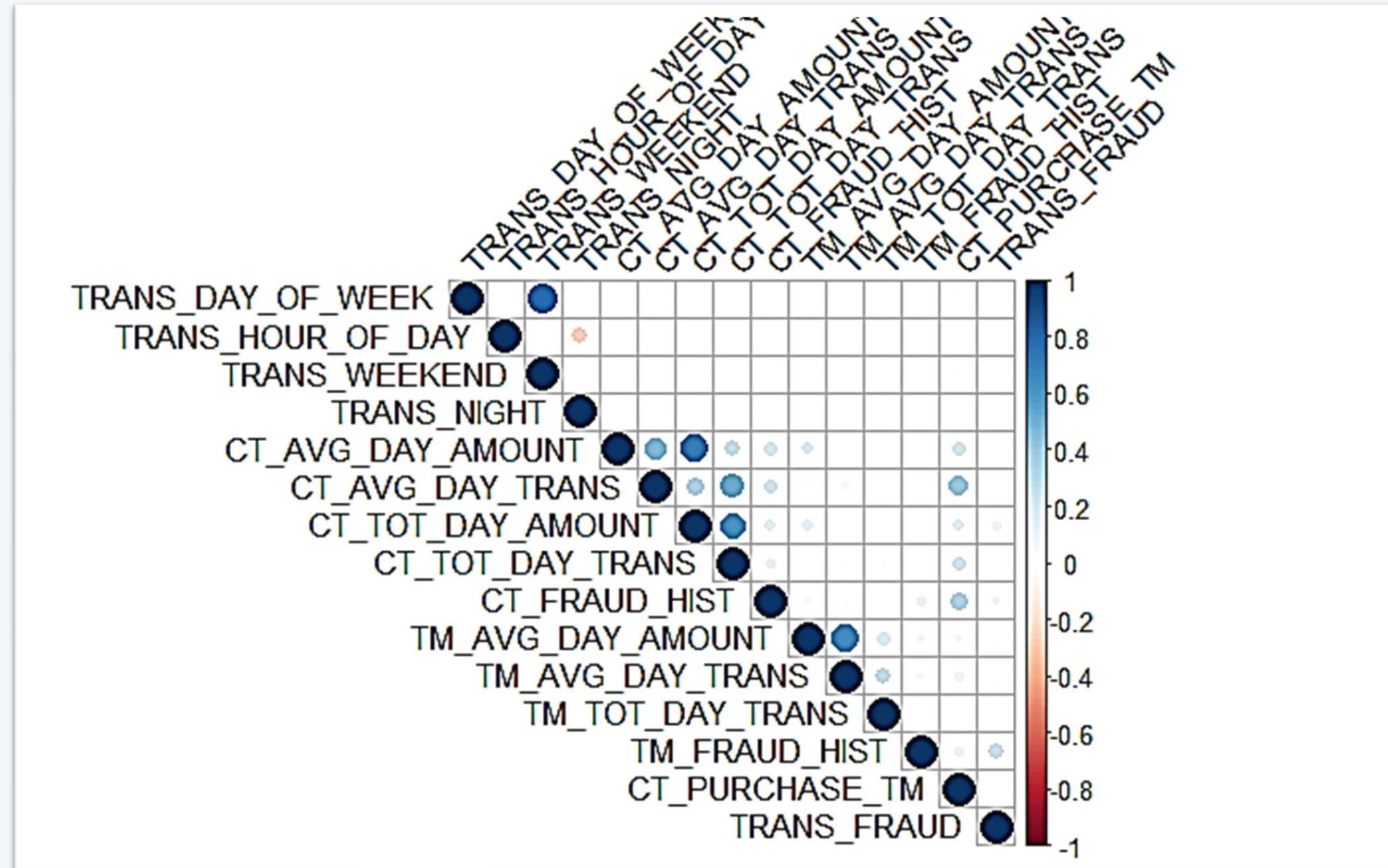
# Explore

---



# Explore

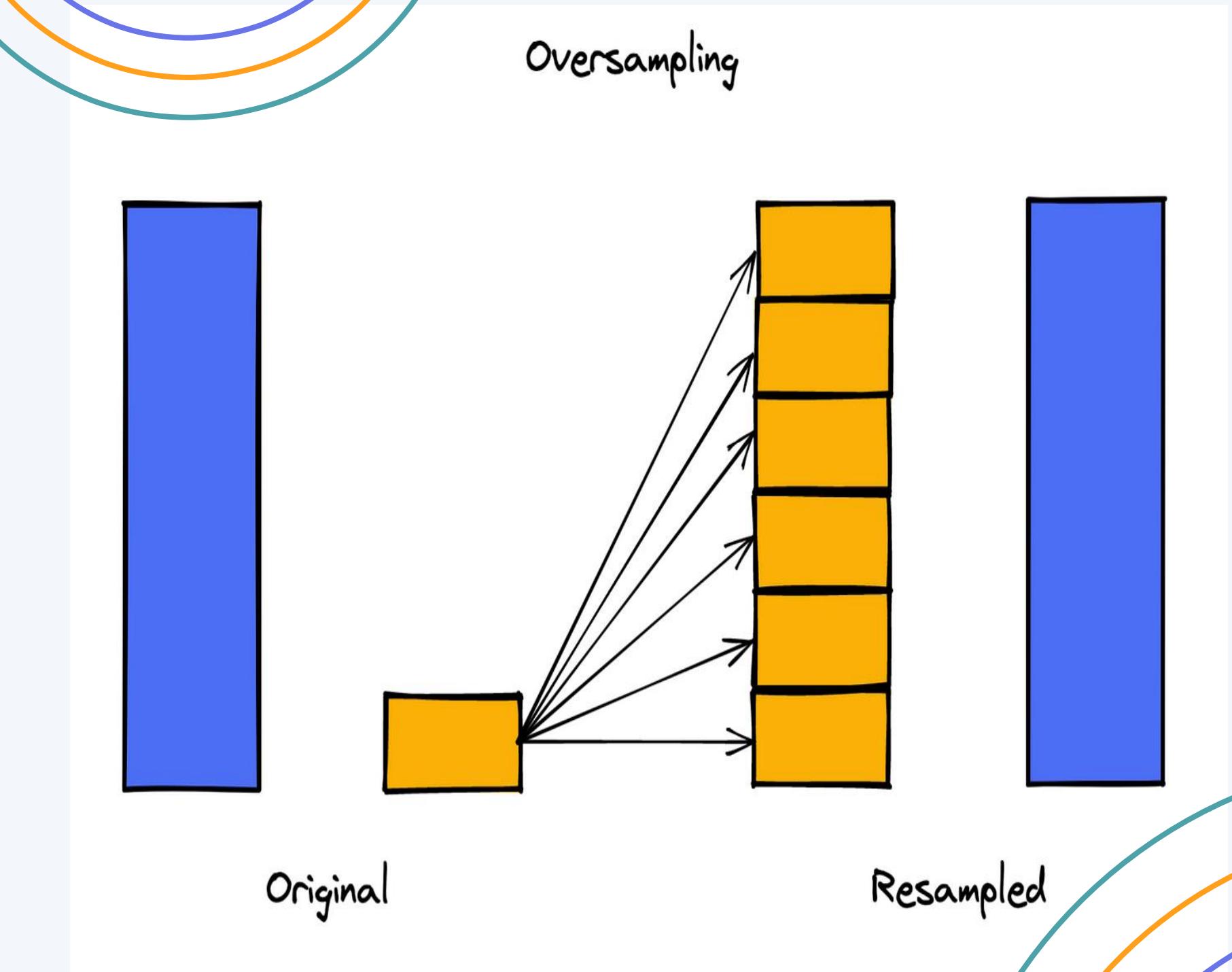
---





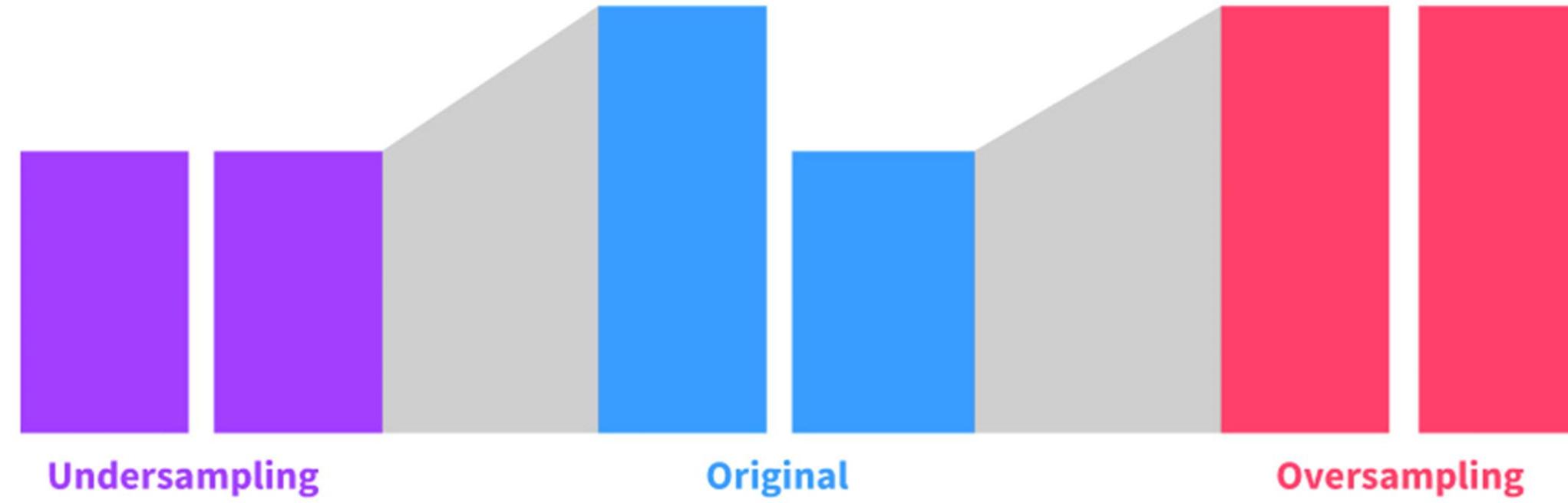
# Resampling

Aleatory process to balance the dataset.



# Resampling

---



In undersampling, we pull all the rare events while pulling a sample of the abundant events in order to equalize the datasets.

Abundant dataset      Rare dataset

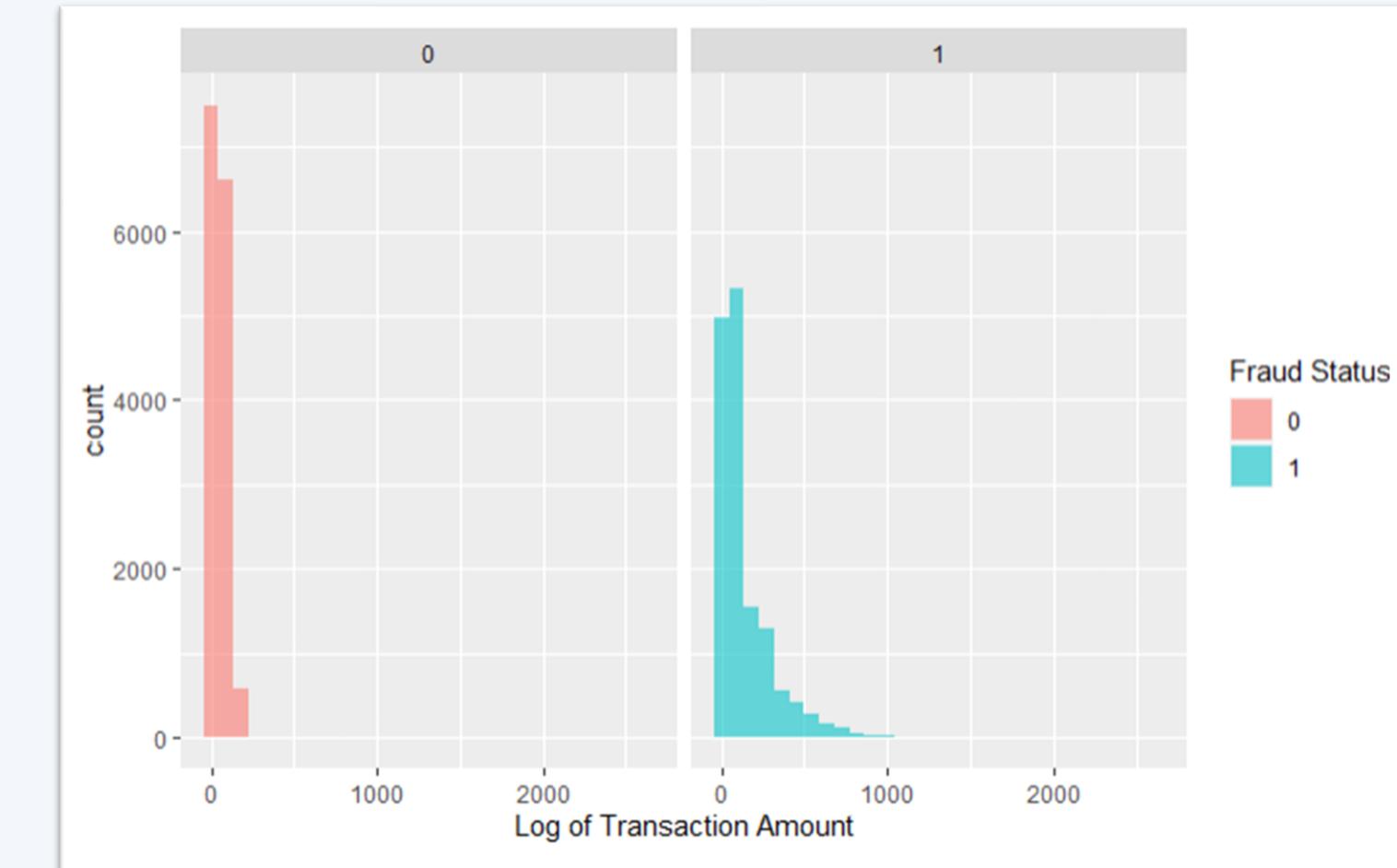
These methods can be used separately or together; one is not better than the other.  
Which method a data scientist uses depends on the dataset and analysis.

# Under-Sampling

---

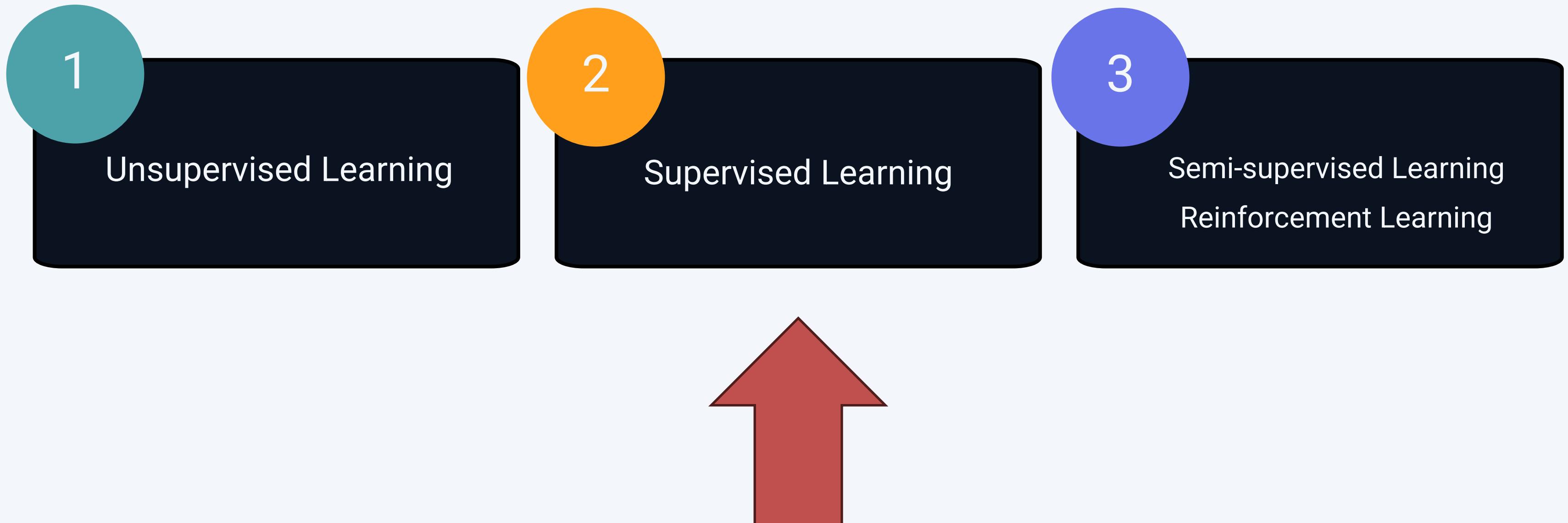


```
tibble [29,362 x 20] (s3: tbl_df/tbl/data.frame)
$ TRANS_FRAUD      : num [1:29362] 1 1 1 1 1 1 1 1 1 1 ...
$ CUSTOMER_ID      : num [1:29362] 0 1 1 1 1 1 1 1 1 1 ...
$ TERMINAL_ID       : num [1:29362] 2352 5 5 1083 1083 ...
$ TRANS_AMOUNT      : num [1:29362] 22.54 205.35 273.6 31.33 0.57 ...
$ TRANS_DAY_OF_WEEK: num [1:29362] 4 7 1 4 5 3 2 1 5 6 ...
$ TRANS_HOUR_OF_DAY: int [1:29362] 12 13 9 10 18 8 12 8 6 18 ...
$ TRANS_WEEKEND     : num [1:29362] 0 1 0 0 0 0 0 0 0 1 ...
$ TRANS_NIGHT       : num [1:29362] 0 0 0 0 0 0 0 0 0 0 ...
$ CT_AVG_DAY_AMOUNT: num [1:29362] 147 191 191 191 191 ...
$ CT_AVG_DAY_TRANS : num [1:29362] 2.4 3.69 3.69 3.69 3.69 ...
$ CT_TOT_DAY_AMOUNT: num [1:29362] 22.5 549.8 321.4 203.8 120 ...
$ CT_TOT_DAY_TRANS : int [1:29362] 1 7 2 5 3 7 5 6 6 6 ...
$ CT_FRAUD_HIST    : num [1:29362] 0 1 10 18 19 20 4 15 16 17 ...
$ TM_AVG_DAY_AMOUNT: num [1:29362] 81.5 91.2 91.2 94 94 ...
$ TM_AVG_DAY_TRANS : num [1:29362] 1.28 1.48 1.48 1.55 1.55 ...
$ TM_TOT_DAY_AMOUNT: num [1:29362] 36.32 324.51 273.6 31.33 0.57 ...
$ TM_TOT_DAY_TRANS : int [1:29362] 2 3 1 1 1 4 4 4 1 2 ...
$ TM_FRAUD_HIST    : num [1:29362] 9 0 1 12 18 21 1 2 13 21 ...
$ TM_CLIENTELE     : int [1:29362] 27 32 32 36 36 36 35 33 33 33 ...
$ CT_PURCHASE_TM   : num [1:29362] 4 1 2 4 5 6 0 2 3 4 ...
```



## Near-Miss Under-Sampling

# Model and Algorithm Selection



# Supervised Algorithms

Logistic Regression



Highly accurate

Decision Trees



Highly reliable

Random Forest



Low computational complexity

Gradient Boosting  
Machine



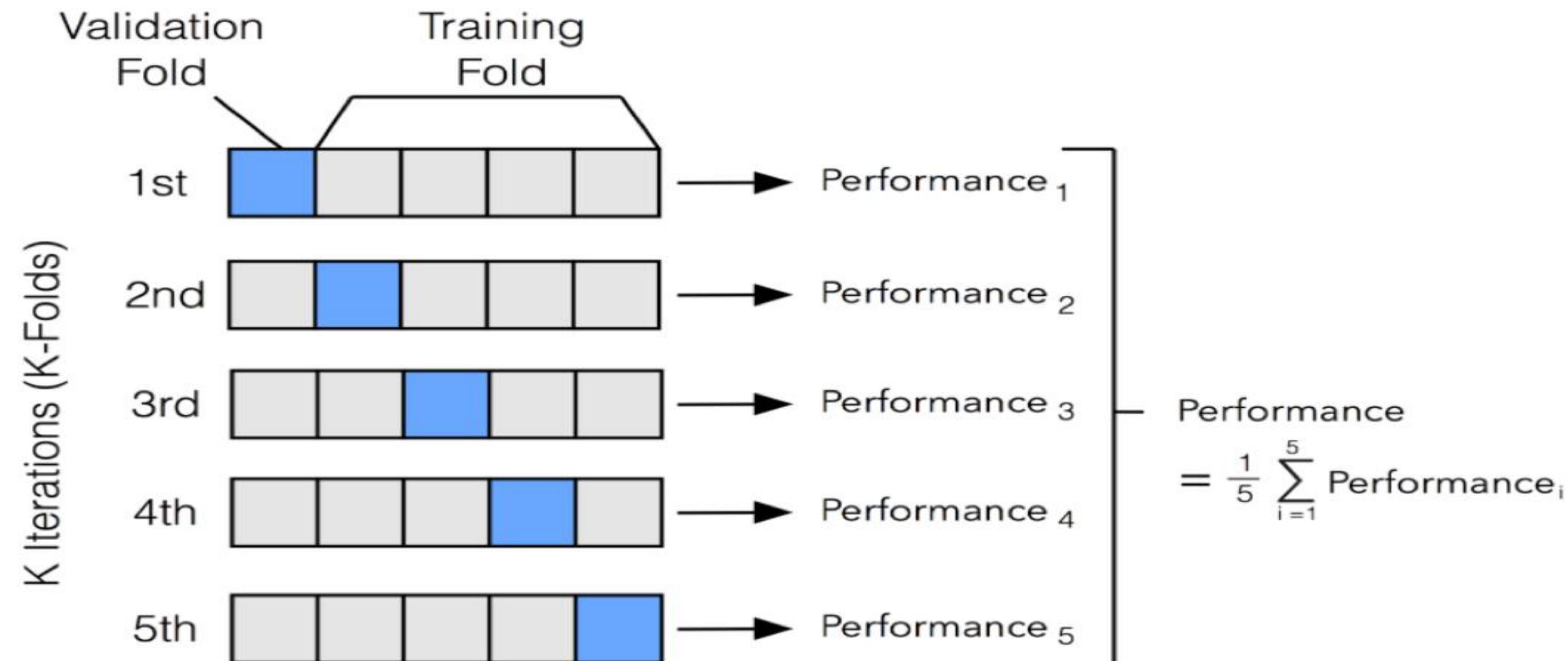
Ease of overfitting

Support Vector  
Machine



Label Dependency

# Cross-Validation



Model should be trained using cross-validation in the training dataset (0.80 split)

# Metrics



Train Dataset

	AUC ROC	Average precision
<b>Logistic regression</b>	0.892	0.663
<b>Decision tree with depth of two</b>	0.802	0.586
<b>Decision tree - unlimited depth</b>	1.000	1.000
<b>Random forest</b>	1.000	1.000
<b>XGBoost</b>	1.000	0.995

Time in Minutes (Balanced Dataset)

	Training execution time	Prediction execution time
<b>Logistic regression</b>	1.566042	0.010171
<b>Decision tree with depth of two</b>	0.079466	0.004938
<b>Decision tree - unlimited depth</b>	0.759311	0.007423
<b>Random forest</b>	2.057255	0.106525
<b>XGBoost</b>	6.960977	0.132526

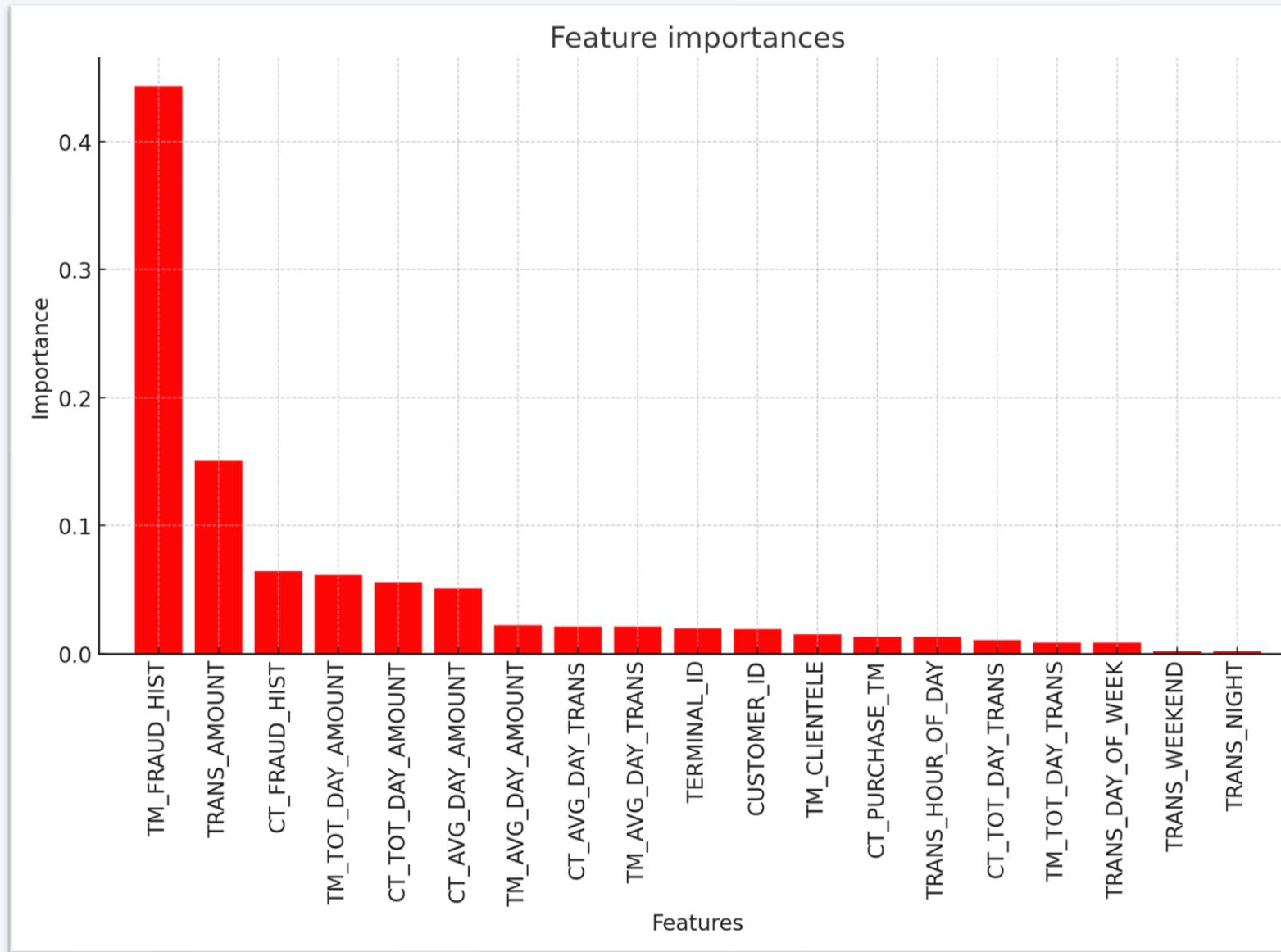
Test Dataset

	AUC ROC	Average precision
<b>Logistic regression</b>	0.871	0.606
<b>Decision tree with depth of two</b>	0.763	0.496
<b>Decision tree - unlimited depth</b>	0.788	0.309
<b>Random forest</b>	0.867	0.658
<b>XGBoost</b>	0.862	0.639

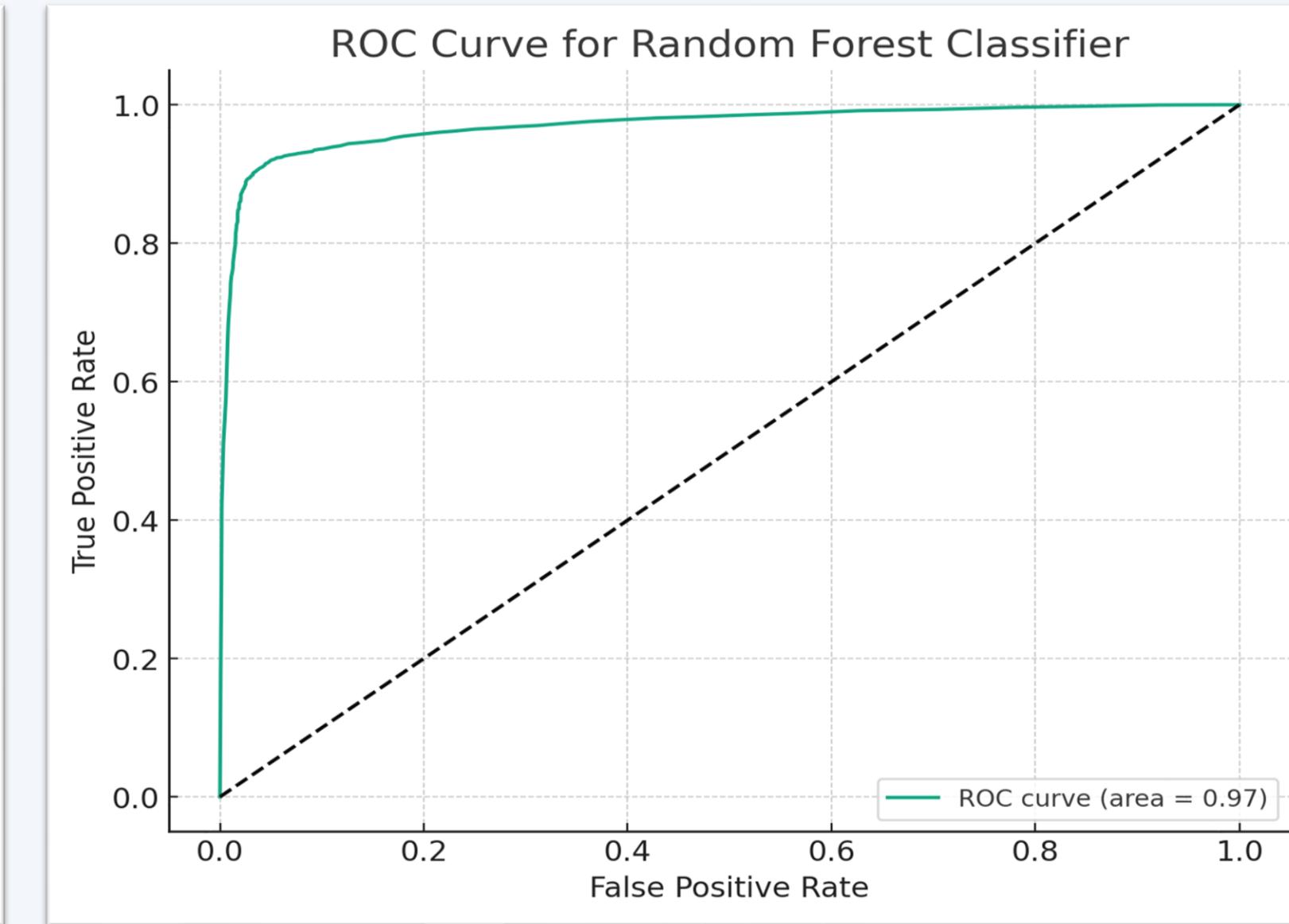
Time in Minutes (Full Dataset)

	Training execution time	Prediction execution time
<b>Logistic regression</b>	28.249504	0.724941
<b>Decision tree with depth of two</b>	5.608281	0.330950
<b>Decision tree - unlimited depth</b>	63.460486	0.684312
<b>Random forest</b>	82.199115	3.443370
<b>XGBoost</b>	107.674864	1.679429

# Metrics

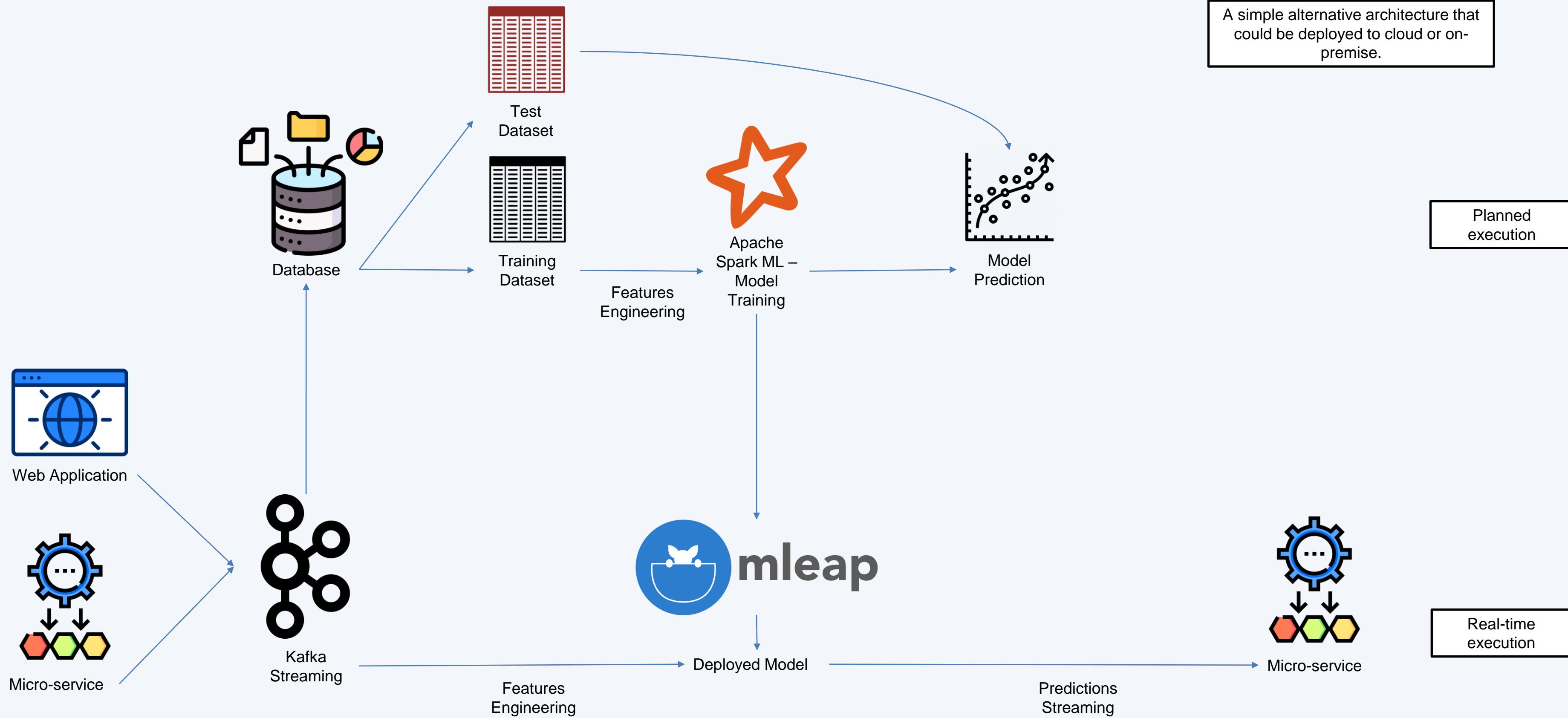


The model confirm the insights coming from the data exploration.



Performance based only on detecting correctly fraudulent transactions.

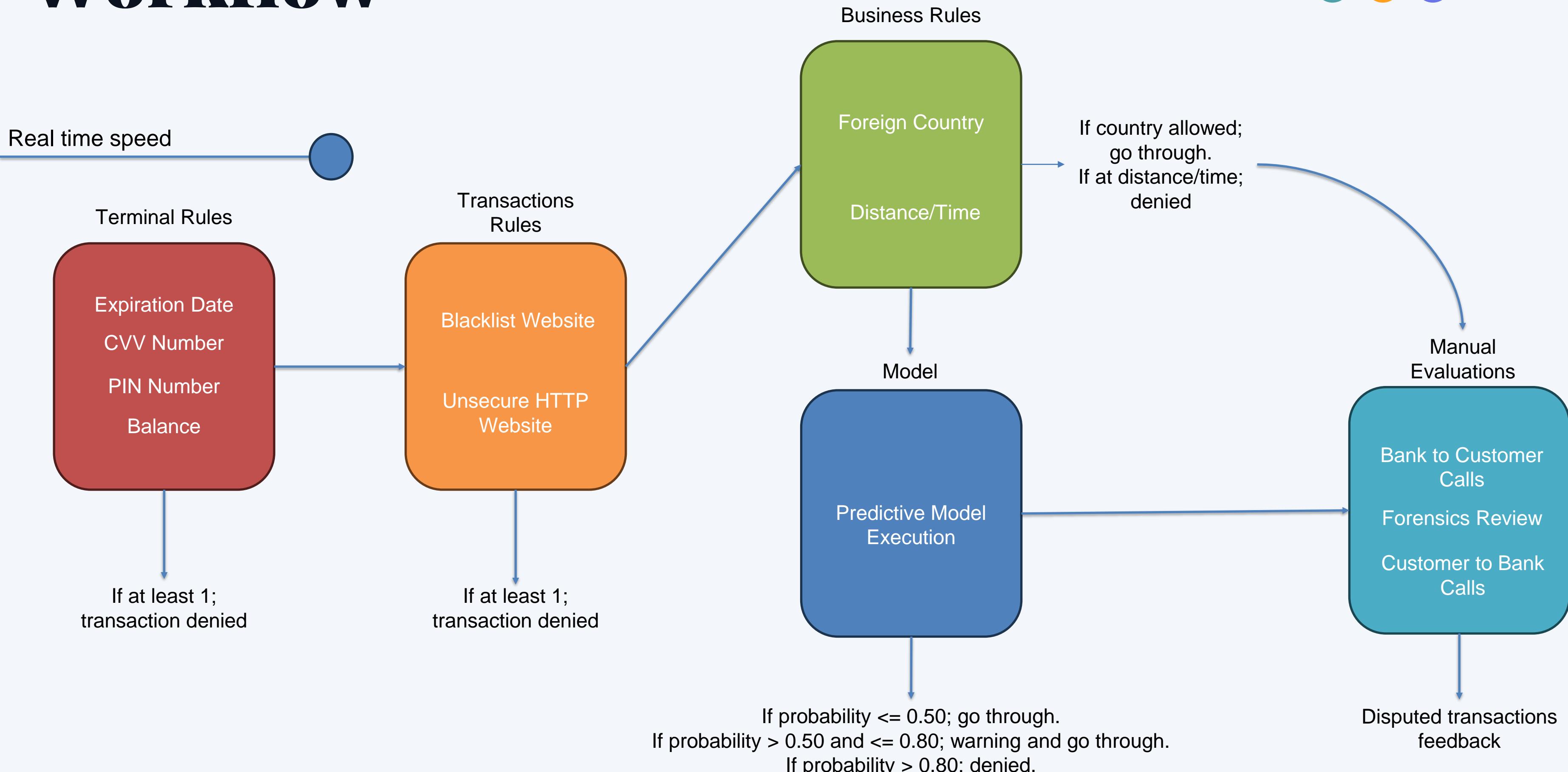
# Architecture



08.

# The System

# Workflow



09.

---

# The Monitor

This Shiny app displays a confusion matrix.

**Choose csv file**

Browse... test(1).csv  
Upload complete

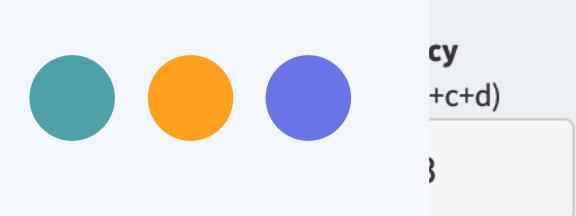
**Reference column**  
reference

**Prediction column**  
prediction

# Monitoring

**Reference:**

		Positive	Negative	Total		
		Positive	Negative	Total	Positive Predictive Value (PPV)	False Discovery Rate (FDR)
Predicted:	Positive	a true positive 54	b false positive 32	a + b 86	a/(a+b) 0.63	1-PPV 0.37
	Negative	c false negative 27	d true negative 231	c + d 258	d/(c+d) 0.9	False Omission Rate (FOR) 1-NPV 0.1
Total	a + c 81	b + d 263	a+b+c+d 344			
	Sensitivity True Positive Rate (TPR) a/(a+c) 0.67	Specificity True Negative Rate (TNR) d/(b+d) 0.88		Prevalence (a+c)/(a+b+c+d) 0.24	Pos Likelihood Ratio (LR+) TPR/FPR 5.48	
					Neg Likelihood Ratio (LR-) FNR/TNR 0.38	



**IT Responsibility:** Keep the system running.

**Business Responsibility:** Keep the system quality through monitoring.

Customer satisfaction could take a hit if False Positive increase.

## Model Performance



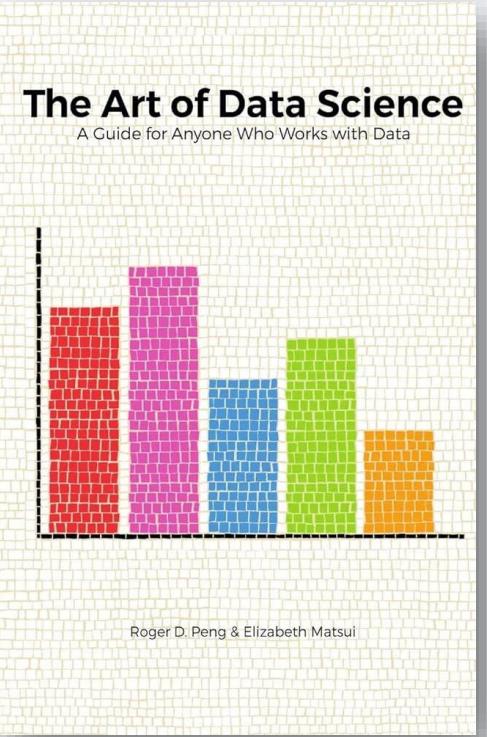
## What was violated

Monitor Name: Loss Rate Monitor  
Monitor Type: Performance Degradation  
Condition: Win Rate < 30%

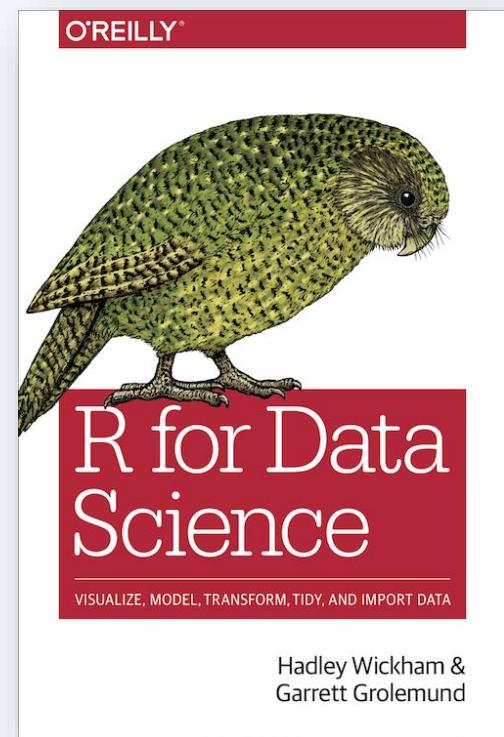
- 1- Every project needs to be planned and KPIs must be established before the kick-off.
- 2- Multidisciplinary teams are quite important; but only the necessary people for the project.
- 3- Data volume is great, but data quality is king.
- 4- Data could discard our 20 years experience hypothesis over a certain business problem.
- 5- Fancy is not always better for real business solutions; pragmatism is key.
- 6- IT are friends, not food.



## **Lessons Learned**



<https://bookdown.org/rdpeng/artofdatascience/>



<https://r4ds.had.co.nz/>

# Google Google Data Analytics Professional Certificate

This is your path to a career in data analytics. In this program, you'll learn in-demand skills that will have you job-ready in less than 6 months. No degree or experience required.

-Taught in English | [Video subtitles available](#)

Instructor: [Google Career Certificates](#)  
[Top Instructor](#)

<https://www.coursera.org/professional-certificates/google-data-analytics>



[https://www.youtube.com/watch?v=V8eKsto3Ug&t=3980s&ab\\_channel=freeCodeCamp.org](https://www.youtube.com/watch?v=V8eKsto3Ug&t=3980s&ab_channel=freeCodeCamp.org)

## Questions?

Guenadie Nibbs  
Data Science | Statistics | Compliance | Risk |  
Software Development  
Dominican Republic - [Contact info](#)

in Barna Management School

<https://www.linkedin.com/in/guenadie-nibbs-1912b933/>



# What's Next Guys?



# thank you!

"Success is no accident. It is hard work, perseverance, learning, studying, sacrifice and most of all, love of what you are doing or learning to do." — Pele

