QUANZETTA

# The Data Has Arrived. Now What?

## Opening your first machine learning dataset

The company has sent us the churn data — where do we start?




churn_data.csv

12.5 KB

| customerID | tenure | Contract | MonthlyCharges | Churn |
|---|---|---|---|---|
| 7590-VHVEG | 1 | Month-to-month | 29.85 | Yes |
| 5575-GNVDE | 34 | One year | 56.95 | No |
| 3668-QPYBK | 2 | Month-to-month | 53.85 | Yes |
| 7795-CFOCW | 45 | Two year | 42.30 | No |
| 9237-HQITU | 8 | Month-to-month | 70.70 | Yes |
| ... | ... | ... | ... | ... |

1 **CSV file**
Raw data from the company

2 **Pandas**
Read the CSV with pandas

3 **DataFrame**
Clean, structured data for analysis

QUANZETTA

# What are we actually looking at?

A dataset is a table of customers and their information.

**One column**
= one type of information

| customerID | tenure (months) | Contract | MonthlyCharges (USD) | Churn |
|---|---|---|---|---|
| 1001 | 1 | Month-to-month | $89.00 | Yes |
| 1002 | 12 | One year | $69.00 | No |
| 1003 | 24 | Two year | $79.00 | No |
| 1004 | 3 | Month-to-month | $99.00 | Yes |
| 1005 | 18 | One year | $59.00 | No |

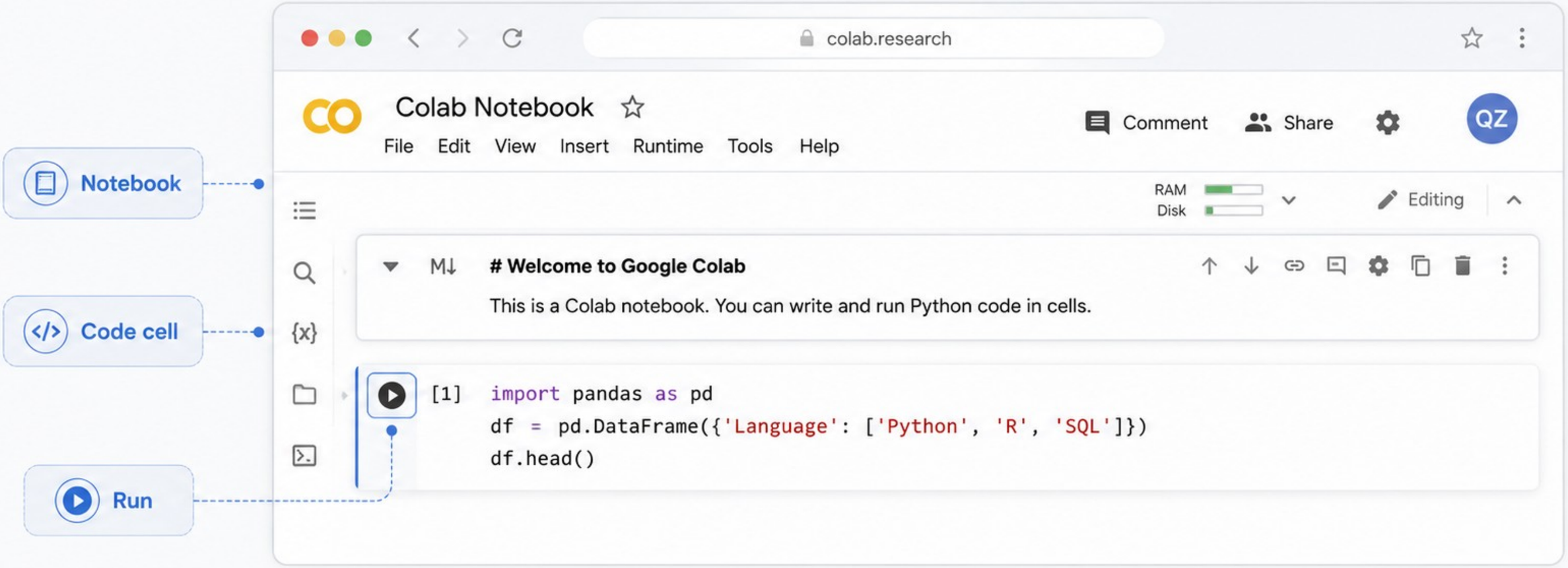
**One row**
= one customer

**Churn column**
= what eventually happened

QUANZETTA

# Open the workspace

## A quick start with Google Colab



QUANZETTA

# Load the dataset for the first time

## Our first code inside a real ML project

Run

```
import pandas as pd

df = pd.read_csv('customer_churn.csv')

df.head()
```

| | customer_id | gender | age | tenure | monthly_charges | churn |
|---|---|---|---|---|---|---|
| 0 | 1001 | Female | 34 | 12 | 70.35 | No |
| 1 | 1002 | Male | 45 | 3 | 89.10 | Yes |
| 2 | 1003 | Female | 28 | 24 | 60.20 | No |
| 3 | 1004 | Male | 52 | 6 | 102.50 | Yes |
| 4 | 1005 | Female | 31 | 1 | 65.00 | No |

**This is your first real data load!**

QUANZETTA

# Meet the DataFrame

## What Pandas gives us after loading the CSV

**CSV file**




sales_data.csv

**Pandas**

pandas

**DataFrame**

| OrderID | Customer | Product | Amount |
|---|---|---|---|
| 1001 | Alice | Laptop | 1200.00 |
| 1002 | Bob | Mouse | 25.50 |
| 1003 | Charlie | Keyboard | 45.00 |
| 1004 | Diana | Monitor | 200.00 |
| 1005 | Ethan | USB Cable | 10.00 |
| ... | ... | ... | ... |

A DataFrame is a table-like structure for working with data in Python.

QUANZETTA

# First look at the customers

## Inspecting the first rows with df.head()

```
df.head()
```

| | customerID | gender | SeniorCitizen | Partner | Dependents | tenure | Contract | MonthlyCharges | TotalCharges | Churn |
|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 7590-VHVEG | Female | 0 | Yes | No | 1 | Month-to-month | 29.85 | 29.85 | No |
| 1 | 5575-GNVDE | Male | 0 | No | No | 34 | One year | 56.95 | 1889.50 | No |
| 2 | 3668-QPYBK | Male | 0 | No | No | 2 | Month-to-month | 53.85 | 108.15 | Yes |
| 3 | 7795-CFOCW | Male | 0 | No | No | 45 | One year | 42.30 | 1840.75 | No |
| 4 | 9237-HQITU | Female | 0 | No | No | 2 | Month-to-month | 70.70 | 151.65 | Yes |

**customerID**
Unique ID assigned to each customer.

**tenure**
Number of months the customer has stayed with us.

**Contract**
Type of contract (Subscribed on).

**MonthlyCharges**
Amount charged to the customer each month.

**Churn**
If the customer has left (Yes / No).

QUANZETTA

# How big is our dataset?

Rows and columns at a glance

```
>>> df.shape
```

```
(7043, 21)
```

**7043**
**rows**
= customers

**21**
**columns**
= information about each customer

```
>>> df.columns
```

```
'customer_id', 'gender', 'senior_citizen', 'partner', 'dependents', 'tenure',
'phone_service', 'multiple_lines', 'internet_service', 'online_security',
'online_backup', 'device_protection', 'tech_support', 'streaming_tv',
'streaming_movies', 'contract', 'paperless_billing', 'payment_method',
'monthly_charges', 'total_charges', 'churn'
```

QUANZETTA

# A quick investigation

## What do you notice about these customers?

| Customer ID | Tenure | Contract | Monthly Charges | Churn |
|---|---|---|---|---|
| 7590-VHVEG | 1 month | Month-to-month | $94.85 | Yes |
| 5575-GNVDE | 12 months | One year | $56.15 | No |
| 3668-QPYBK | 24 months | Two year | $19.90 | No |
| 7795-CFOCW | 2 months | Month-to-month | $104.70 | Yes |

Short tenure?

Monthly contract?

Higher charges?

**What patterns do you notice?**

QUANZETTA

# Not all columns are the same

## Can Machine Learning use them as they are?

### tenure
→ number

| tenure |
|---|
| 1 |
| 5 |
| 12 |
| 24 |
| 36 |
| ... |

**Number**
Quantitative values we can measure.

### Contract
→ category

| Contract |
|---|
| Month-to-month |
| One year |
| Two year |
| Month-to-month |
| One year |
| ... |

**Category**
Groups or labels with no order.

### customerID
→ identifier

| customerID |
|---|
| 7590-VHVEG |
| 5575-GNVDE |
| 3668-QPYBK |
| 7795-CFOCW |
| 9237-HQITU |
| ... |

**Identifier**
Unique IDs that only identify each row.

### Churn
→ yes/no outcome

| Churn |
|---|
| No |
| Yes |
| No |
| No |
| Yes |
| ... |

**Yes/No Outcome**
The target we want to predict.

**Next lesson: Understanding Your Data**