

1) Charger les données

Fichier Excel (.xlsx) ou CSV

Drag and drop file here

Limit 200MB per file • XLSX, XLS, CSV

Browse files



Base de données plu...
18.5KB

2) Nettoyage / Outliers

☒ Supprimer les outliers (IQR par complexité)

Facteur IQR (k)

1.50

(L'explication détaillée du nettoyage est affichée dans la page pour être plus visible.)

PLUTO – Module formatif : analyse de la durée (mois)

Aperçu des données brutes

	Line		complexity	effort mxy	duration months	revisions number
10	11	follower		1.081	25.0645	
11	12	Innovative		2.2494	26.6129	
12	13	1st in family		1.9636	87.1613	
13	14	follower		0.1496	26.6129	
14	15	follower		0.3298	26.9032	
15	16	follower		0.4671	28.5806	
16	17	follower		1.1934	23.2258	
17	18	follower		3.429	16.9355	
18	19	Innovative		7.1364	28.3068	
19	20	Innovative		10.8667	49.0323	

Dimensions: 178 lignes × 6 colonnes

Étape 1A — Nettoyage : suppression des valeurs manquantes (NA)

Lignes initiales

178

Lignes supprimées (NA)

0

% supprimé (NA)

0.00%

Les lignes avec NA sur les colonnes clés ne permettent pas de calculer correctement les indicateurs, ni d'ajuster un modèle prédictif.

Liste des lignes supprimées (NA)

Aucune ligne supprimée pour NA.

Données après nettoyage NA (base de travail)

	complexity	duration_months	effort_my	revisions_number
10	follower	25.0645	1.081	
11	innovative	26.6129	2.2494	
12	first_in_family	87.1613	1.9636	
13	follower	26.6129	0.1496	
14	follower	26.9032	0.3298	
15	follower	28.5806	0.4671	
16	follower	23.2258	1.1934	
17	follower	16.9355	3.429	
18	innovative	28.3068	7.1364	
19	innovative	49.0323	10.8667	

Dimensions après NA: 178 lignes × 5 colonnes

Étape 1B — Valeurs extrêmes (outliers) sur Duration

Objectif : obtenir une base représentative des projets standards (sans cas extrêmes rares).

 Nettoyage & outliers (explication)

On commence par supprimer les lignes avec des valeurs manquantes (NA) sur les colonnes clés, car elles empêchent de calculer correctement les statistiques et d'ajuster un modèle.

Ensuite, si l'option est activée, on détecte les **valeurs atypiques** (outliers) de *Duration* avec la règle **IQR** calculée **par complexité** :

- $Q1 = 25\text{e percentile}$, $Q3 = 75\text{e percentile}$, $IQR = Q3 - Q1$
- bornes : $Q1 - k \times IQR$ et $Q3 + k \times IQR$

Une durée est un outlier si elle sort de ces bornes. Le facteur k (1.5 par défaut) contrôle la sévérité.

Lignes avant outliers

178

Outliers supprimés

4

% supprimé (outliers)

2.25%

Résumé outliers (bornes IQR par complexité)

	complexity	total	outliers	pct_outliers	lower
0	first_in_family	27	1	3.7	
1	follower	81	3	3.7	
2	innovative	70	0	0	

Liste des lignes supprimées (outliers)

	complexity	duration_months	effort_my	revisions_number
0	first_in_family	87.1613	1.9636	
1	follower	63.9677	0.7027	
2	follower	72.2258	0.3477	
3	follower	63.1935	2.0211	

Étape 2 — Statistiques descriptives (Duration)

On résume la variable *Duration* (globalement et par complexité) avec des statistiques descriptives. Ces indicateurs servent de base pour interpréter les distributions et calibrer les bornes des analyses.

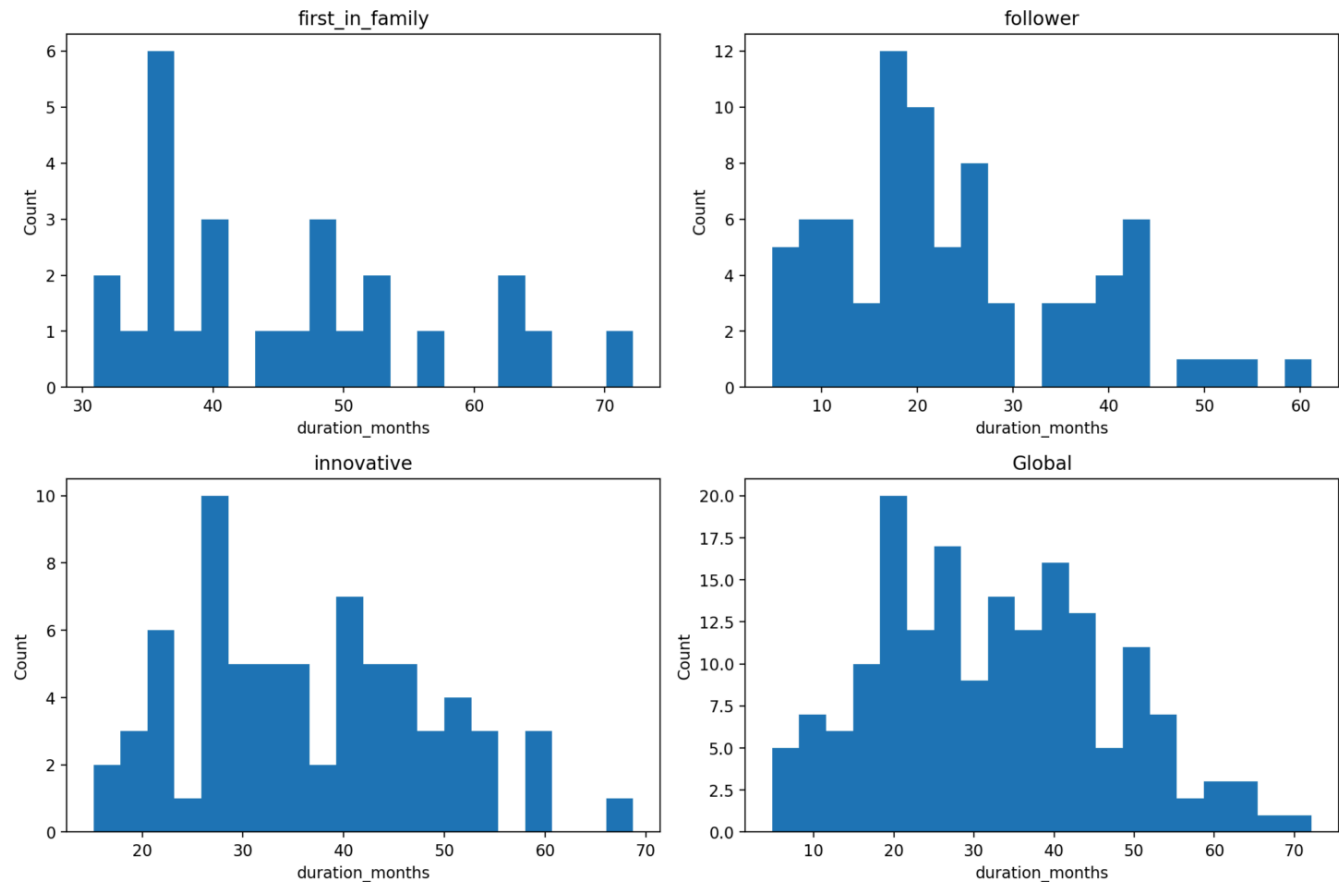
Statistiques descriptives — global

	N	Min	Max	Moyenne	Médiane	Écart-type (STD)	Erreur standard (SE)
0	174	4.8060	72.1640	32.3230	31.9840	14.6230	1.1090

Statistiques descriptives — par complexité

	complexity	N	Min	Max	Moyenne	Médiane	Écart-type (s
0	first_in_family	26	30.839	72.164	45.763	43.113	
1	follower	78	4.806	61.194	23.94	20.935	
2	innovative	70	15.161	68.742	36.672	34.645	

Distributions (Duration) — global & par complexité



Étape 3 — Test de normalité (par complexité)

On teste si la durée suit une distribution normale pour chaque niveau de complexité. Ce test sert à orienter le choix du modèle statistique.

🔍 Pourquoi Shapiro-Wilk ?

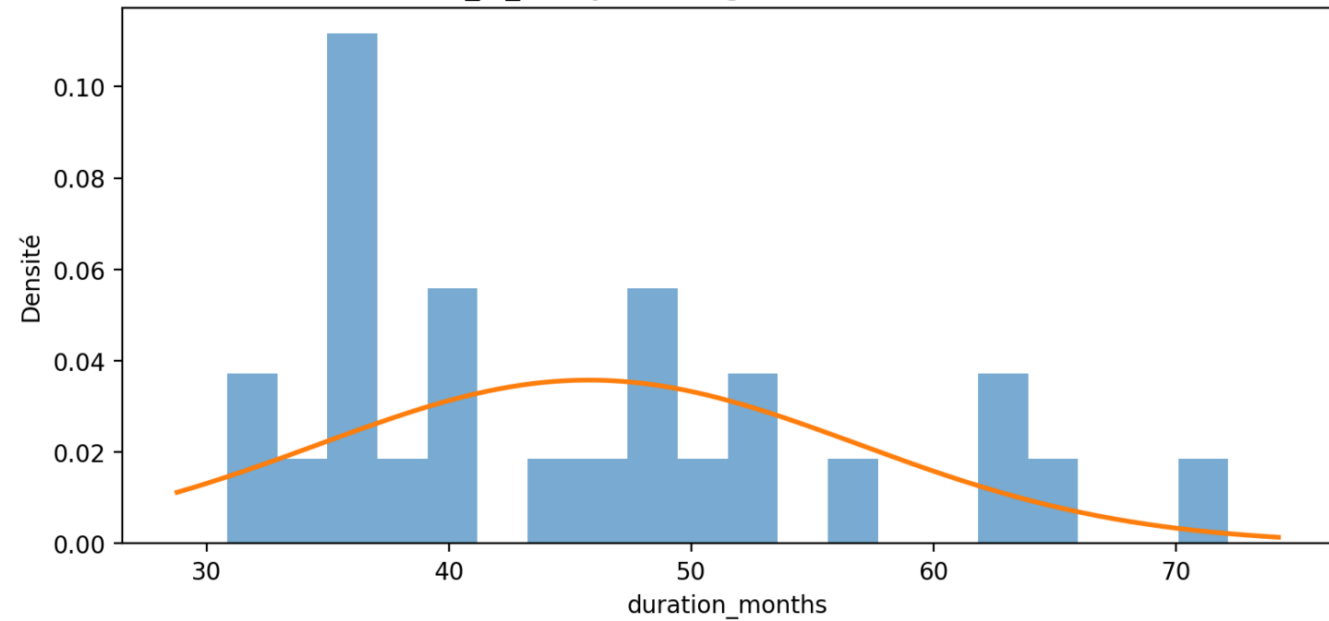
Le test de Shapiro-Wilk vérifie l'hypothèse nulle **H0 : la variable suit une loi normale**. Il renvoie une *p-value* : si $p \geq 0,05$, on ne rejette pas H0 (normalité plausible) ; si $p < 0,05$, on rejette H0 (normalité non plausible). Dans PLUTO, ce diagnostic sert à justifier le recours à des lois alternatives (lognormale, gamma, Weibull, etc.).

Résultats du test de normalité (Shapiro-Wilk)

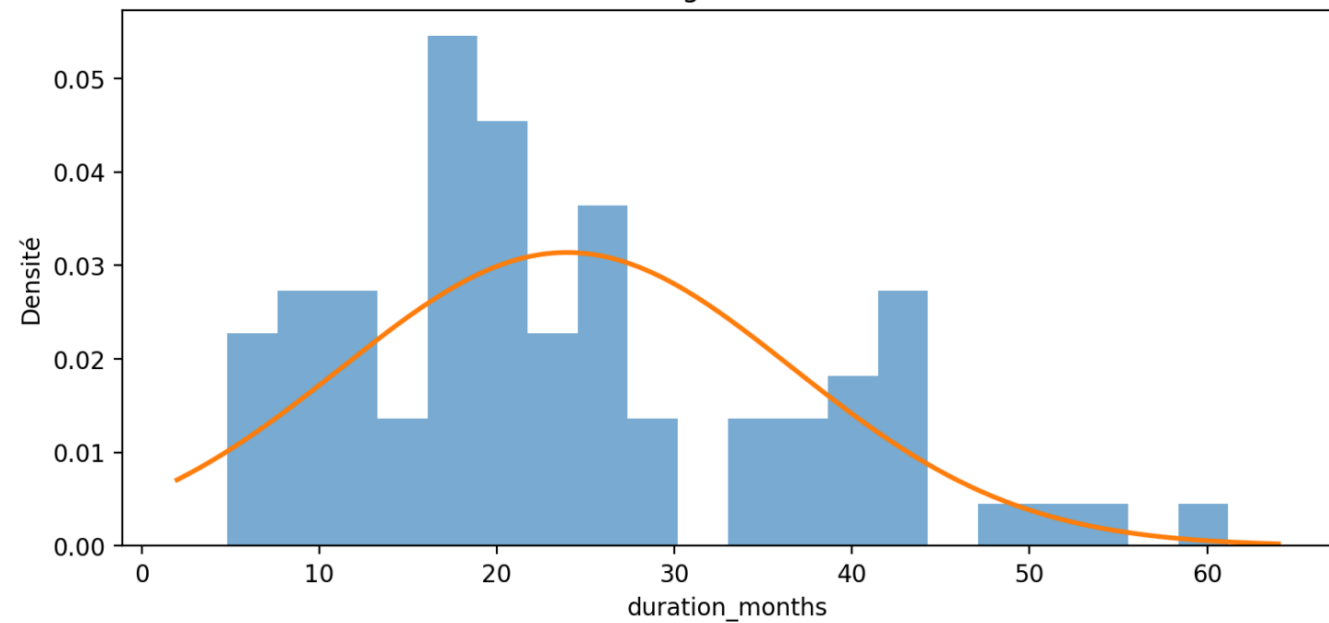
	complexity	n	mean	median	std	shapiro_p
0	first_in_family	26	45.76	43.11	11.15	
1	follower	78	23.94	20.94	12.71	
2	innovative	70	36.67	34.65	12	

Visualisation : histogramme + loi normale

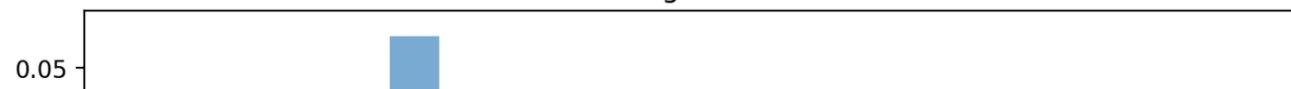
first_in_family — histogramme + loi normale

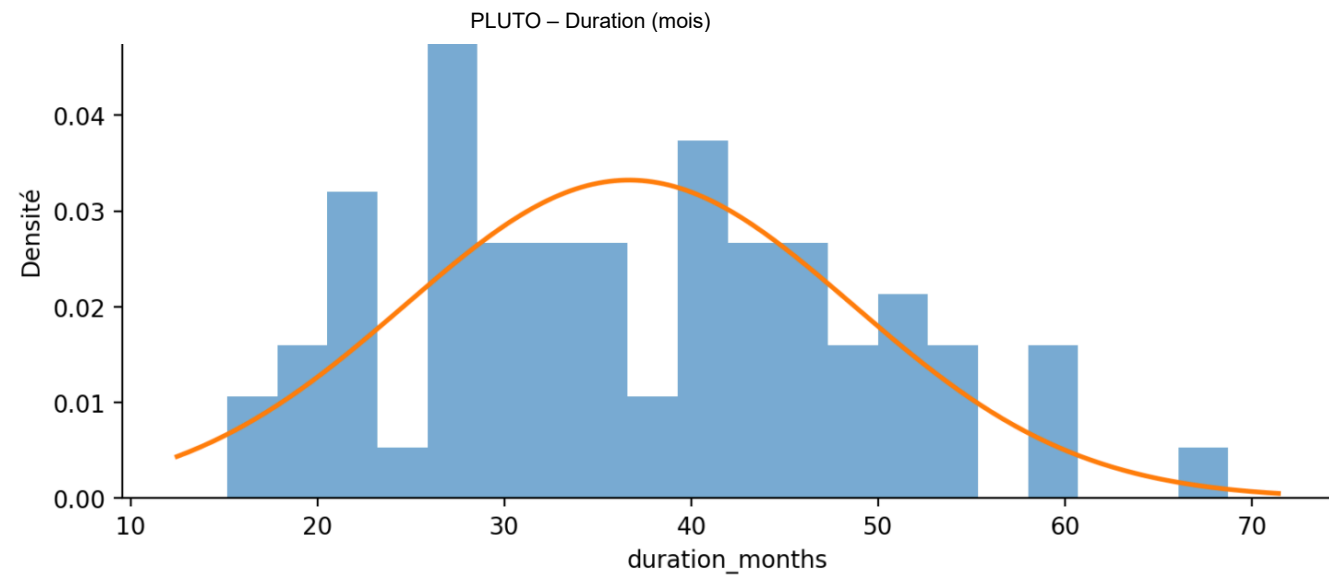


follower — histogramme + loi normale



innovative — histogramme + loi normale





Étape 3.5 — Ajustement des lois et comparaison (AIC / BIC)

On ajuste plusieurs lois candidates sur *Duration* par complexité (par MLE), puis on compare AIC/BIC. Ensuite, on peut visualiser les courbes ajustées sur l'histogramme pour comprendre le choix du meilleur modèle.

Tableau comparatif des modèles (par complexité)

	complexity	model	n	aic
1	first_in_family	lognorm	26	
0	first_in_family	normal	26	
2	first_in_family	gamma	26	
3	first_in_family	weibull	26	
4	first_in_family	expon	26	
7	follower	gamma	78	
8	follower	weibull	78	
6	follower	lognorm	78	
5	follower	normal	78	
9	follower	expon	78	

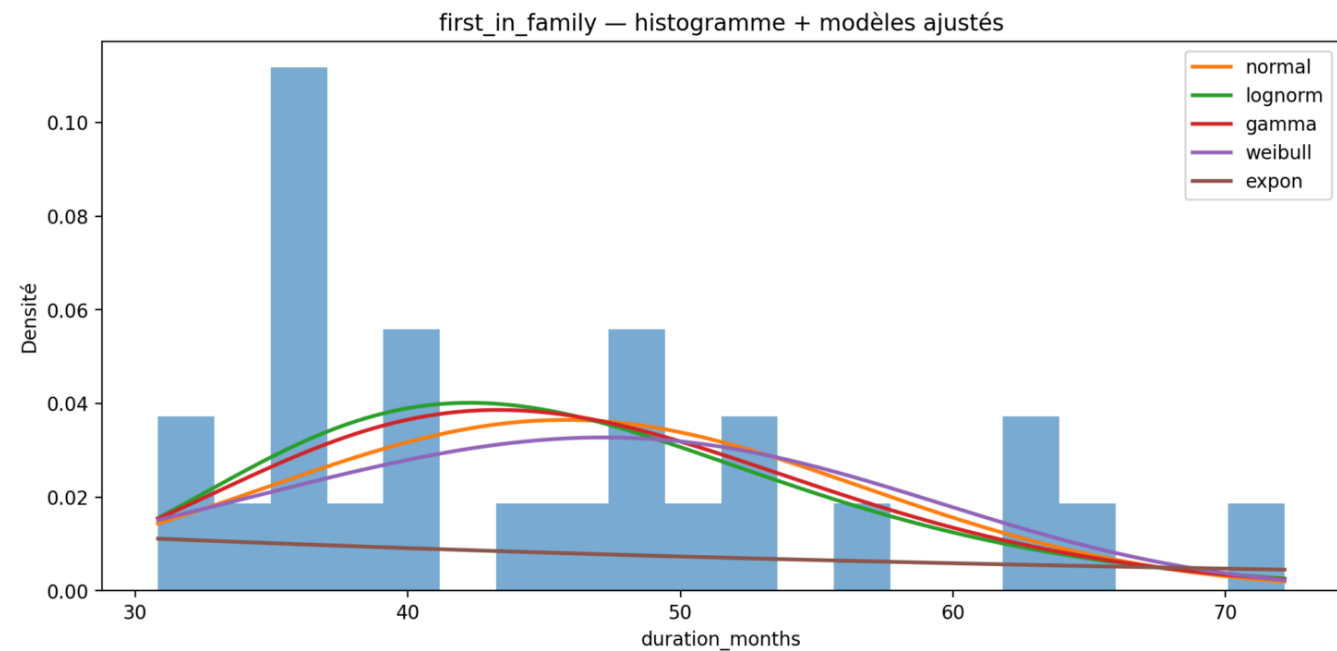
Meilleur modèle par complexité (critère BIC)

	complexity	model	n	aic
0	first_in_family	lognorm		26
1	follower	gamma		78
2	innovative	normal		70

Visualisation : histogramme + modèles ajustés

Choisir une complexité (visualisation)

first_in_family



Étape 4 — Exercice étudiant : sélectionner la meilleure loi

À partir des résultats AIC/BIC (Étape 3.5), l'étudiant propose une loi par complexité. PLUTO compare ensuite ce choix au meilleur modèle (selon BIC) et explique l'écart. On ne ré-affiche pas les courbes ici : elles sont déjà visibles en fin d'étape 3.5.

Choix étudiant

Choisis une distribution pour la complexité : first_in_family

normal



Choisis une distribution pour la complexité : follower

normal



Choisis une distribution pour la complexité : innovative

normal



Feedback PLUTO (comparaison BIC)

Complexité : first_in_family

Ton choix : normal

PLUTO (meilleur BIC) : lognorm

Verdict :  Acceptable

Ton choix (**normal**) n'est pas le meilleur, mais il est **très proche** du meilleur selon BIC ($\Delta BIC = 0.34$).

PLUTO sélectionne **lognorm** car il minimise BIC pour la complexité **first_in_family** (meilleur BIC = 204.34 vs ton BIC = 204.68).

Complexité : follower

Ton choix : normal

PLUTO (meilleur BIC) : gamma

Verdict :  Non optimal

Ton choix (**normal**) est **nettement moins adapté** que le meilleur selon BIC ($\Delta BIC = 8.80$).

PLUTO sélectionne **gamma** car il minimise BIC pour la complexité **follower** (meilleur BIC = 616.90 vs ton BIC = 625.70).

Complexité : innovative

Ton choix : normal

PLUTO (meilleur BIC) : normal

Verdict :  CorrectTon choix (**normal**) est le meilleur selon BIC.PLUTO sélectionne **normal** car il minimise BIC pour la complexité **innovative** (meilleur BIC = 554.07 vs ton BIC = 554.07).

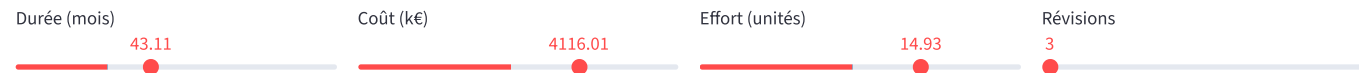
Étape 5 — Hyperplan par complexité : Mahalanobis + percentile (succès)

Ici, on montre la puissance de PLUTO : au lieu de regarder *Duration* seule, on place le projet dans un espace multivarié (hyperplan) **par complexité** avec les variables du cœur : durée, coût, effort, révisions. On calcule ensuite une distance à la distribution historique et on transforme cette distance en **percentile** pour obtenir un score de succès (projet plus ou moins 'typique' et donc plus ou moins réaliste).

Complexité (hyperplan)

first_in_family

Curseurs étudiant (nouveau projet)



Cadre statistique et mathématique (pas à pas)

On travaille par **complexité** : chaque complexité a son propre espace (hyperplan). Pour cette complexité :

1. on prend les variables du cœur : durée, coût, effort, révisions ;
2. on applique une normalisation **z-score** (par complexité) ;
3. on estime la covariance Σ du nuage (sur les z-scores) ;
4. on calcule la distance de Mahalanobis du projet utilisateur ;
5. on convertit cette distance en percentile empirique (ECDF) pour obtenir un score de succès.

$$z = \frac{x - \mu_c}{\sigma_c}$$

$$\Sigma_c = \text{Cov}(Z) \quad (\text{estimée sur l'historique})$$

$$d(z) = \sqrt{z^T \Sigma_c^{-1} z}$$

$$\text{Succès} = 100 \times (1 - \hat{F}_c(d(z)))$$

Résultat : distance et score

Distance de Mahalanobis

2.353

Percentile ECDF

84.6%

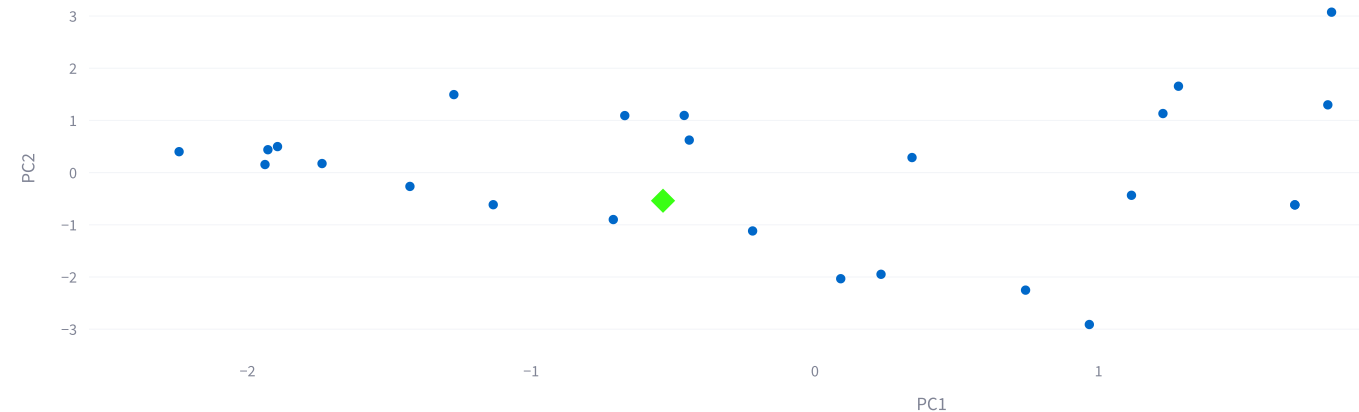
Score de succès

15.4%

Interprétation : si ton point est **central** dans le nuage historique de cette complexité (distance faible), la proportion de projets 'plus proches' que toi est faible, et donc le **score** $100 \times (1 - \text{ECDF})$ est élevé.

Visualisation de l'hyperplan (projection 2D par PCA)

Hyperplan (PCA) — complexité : first_in_family



Validation : pourquoi cette probabilité de succès a du sens

Je le justifie simplement : on compare ton projet aux projets historiques **de la même complexité**. Comme les variables n'ont pas la même unité (mois, k€, effort, révisions), on commence par les mettre sur une même échelle avec le z-score. Sinon, une variable « grosse » écrase toutes les autres.

Ensuite, au lieu de mesurer une distance « naïve » (euclidienne), on utilise Mahalanobis parce qu'elle prend en compte **les corrélations** entre variables. Exemple : si coût et durée bougent souvent ensemble, la distance Mahalanobis ne va pas te pénaliser deux fois pour la même information.

Le point important, c'est que dans un nuage multivarié, les points « typiques » sont au centre, et les points atypiques sont loin. La distance de Mahalanobis sert justement à dire si ton point est plutôt central ou plutôt en dehors de la zone habituelle pour cette complexité.

Enfin, on transforme la distance en **percentile empirique (ECDF)** : on regarde où se situe ta distance par rapport aux distances des projets historiques. C'est volontairement empirique : on n'est pas obligé de supposer que tout est parfaitement gaussien. On utilise les données telles qu'elles sont.

Donc la « probabilité de succès » ici se lit comme une probabilité de **réalisme** : si ton point ressemble à la majorité des projets historiques (distance faible), le score est élevé. Si ton point est très loin (distance forte), le score baisse parce que ce scénario est moins fréquent dans l'historique.

Remarque : si on supposait une normalité multivariée parfaite, $d(z)^2$ suivrait une loi du chi-deux. Ici on préfère l'ECDF, plus robuste et plus simple à expliquer.