

From Flat to Mountain

A Data-Driven Look at Rider Performance

TU Dortmund University
Summersemester 2026

Hossein Mirzagol
November 15, 2025

Contents

1	Detailed Problem Description	1
2	Methods	1
2.1	Data handling and scaling	1
2.2	Descriptive statistics and graphics	2
2.3	Two-way ANOVA and effect sizes	2
2.4	Robust inference	2
2.5	Intraclass correlation coefficient	3
2.6	Mixed-effects modeling	3
2.7	Estimated marginal means and contrasts	3
2.8	Bootstrap confidence intervals for effect sizes	4
2.9	Residual diagnostics	4
2.10	Likelihood diagnostics and optimizer checks	4
2.11	Stage-specific Kruskal–Wallis tests	4
2.12	Zero-inflated Poisson regression (future work)	5
3	Evaluation	5
3.1	Key figures	5
3.2	Hypothesis testing and modeling	7
3.3	Mixed-effects results	7
3.4	Estimated marginal means	7
3.5	Diagnostic interpretation	8
3.6	Estimated marginal means and contrasts	8
3.7	Diagnostics	9

4	Summary and Recommendations	9
----------	------------------------------------	----------

5	Bibliography	9
----------	---------------------	----------

1 Detailed Problem Description

The race organizer delivered a single whitespace-delimited export with quoted rider names, five variables (rider name, rider class, stage, points, stage class), and no missing values[1]. Stage classes were normalized to `flat/hills/mount` according to the organizer’s course taxonomy[1], rider classes were treated as unordered categories, and the response variable is the integer number of points awarded on that stage. Because a single rider appears on multiple stages, observations are nested within riders and therefore constitute clustered data that merit hierarchical treatment[3]. The essential business questions are: (i) do rider classes differ materially in typical point totals, (ii) does stage profile amplify or mute those differences, and (iii) after accounting for repeated measures, do the substantive rankings change? These questions drive the methods in the next section and the evidence reported later.

2 Methods

All computations were run in Python 3.10 using `pandas` [4] for data handling, `seaborn` [5] and `matplotlib` for graphics, `scipy` [6] for distributional tests, and `statsmodels` [7] for modeling. The notebook `cycling.ipynb` is the executable record; this section explains every analytical building block used there.

2.1 Data handling and scaling

Whitespace-separated records were parsed with explicit headers using `pandas.read_csv` and related categorical utilities[4]. Categorical variables were stored as ordered `category` dtypes, preserving the intended stage order (`flat → hills → mount`) and ranking rider classes by their mean points. Point totals remain integers; no standardization is applied because the units are meaningful and comparable across stages. Sanity checks (row counts, rider counts, stage counts) guard against accidental filtering and are reprinted at the start of the notebook to ensure reproducibility.

2.2 Descriptive statistics and graphics

We report conventional summaries (mean, median, variance, standard deviation, interquartile range, skewness) at the overall level and within each rider/stage slice, following standard exploratory data analysis practice[8, 9]. Histograms are defined as frequency estimators over equally spaced bins—formally $h(b) = \frac{1}{n\Delta} \sum_{i=1}^n \mathbf{1}\{x_i \in [b, b + \Delta)\}$ —and reveal where point masses concentrate, while boxplots show medians, quartiles, and whiskers extending to 1.5 times the interquartile range. These tools establish how skewed and zero-inflated the distributions are before deploying heavier statistical machinery.

2.3 Two-way ANOVA and effect sizes

To quantify how much rider class and stage class explain point totals, we fit

$$\text{extpoints} = \beta_0 + \beta_{\text{class}} + \beta_{\text{stage}} + \beta_{\text{class} \times \text{stage}} + \varepsilon,$$

where factor effects are coded with treatment contrasts[10]. Type-II sums of squares provide F -tests that respect marginality in unbalanced designs[11]. Effect sizes use eta-squared η^2 (proportion of total variation explained) and partial eta-squared (proportion of effect-plus-residual variation explained), interpreted per[12]. Because eta-squared can be optimistic, we bootstrap 300 resamples to obtain percentile confidence intervals, following the guidance in[10].

2.4 Robust inference

Classical ANOVA assumes normally distributed, homoskedastic residuals. To check these assumptions we employ the Shapiro–Wilk test (sensitive to heavy tails)[13], Levene’s test (median-centered version to detect unequal variances)[14], and QQ/residual plots (Figures 2 and 3). When assumptions fail, we report (i) Kruskal–Wallis rank tests within each stage to compare class medians[15] and (ii) HC3 heteroskedasticity-consistent Wald tests[16, 17], which rescale the covariance matrix to remain valid under variance heterogeneity.

2.5 Intraclass correlation coefficient

Within the mixed-effects framework, the intraclass correlation coefficient (ICC) quantifies the proportion of unexplained variation attributable to rider identity[3].

$$ICC = \frac{\sigma_u^2}{\sigma_u^2 + \sigma_\varepsilon^2},$$

where σ_u^2 is the random-intercept variance and σ_ε^2 the residual variance. Higher ICC values indicate stronger clustering of repeated rider measurements and justify hierarchical modeling.

2.6 Mixed-effects modeling

Because each rider appears multiple times, we fit a random-intercept mixed linear model, often written as

$$\text{extpoints}_{ij} = (\beta_0 + u_i) + (\beta_{\text{class}} + \beta_{\text{stage}} + \beta_{\text{class}\times\text{stage}}) + \varepsilon_{ij},$$

where $u_i \sim \mathcal{N}(0, \sigma_u^2)$ captures rider-specific baselines[3]. This accounts for within-rider correlation. The intraclass correlation coefficient (ICC) reports the proportion of unexplained variance attributable to rider identity[3]. Mixed models guard against inflated Type I error when repeated measures are present.

2.7 Estimated marginal means and contrasts

We compute estimated marginal means (EMMs) from the fitted ANOVA design matrix: each rider-class/stage-class pair is projected onto the coefficient space to obtain "balanced" means, along with standard errors and t -based confidence intervals[18]. Pairwise contrasts are built as linear combinations of coefficients; Holm adjustment controls the family-wise error rate across multiple comparisons[19]. Reporting EMMs complements raw means because it adjusts for unbalanced counts.

2.8 Bootstrap confidence intervals for effect sizes

To complement point estimates of partial eta-squared (η_p^2) we draw $B = 300$ bootstrap resamples with replacement from the rider-stage rows, recomputing the two-way ANOVA on each sample. The percentile interval is then reported as $[q_{0.025}, q_{0.975}]$, where q_α denotes the α quantile of the bootstrap distribution[10, 12]. Because the procedure is distribution-free, it supplies uncertainty measures even when ANOVA assumptions are stressed or violated.

2.9 Residual diagnostics

Model adequacy is checked through QQ plots that compare standardized residuals to the theoretical normal quantiles and through residual-versus-fitted plots that reveal heteroskedasticity patterns[9, 8]. Fan-shaped residual spreads prompt the use of HC3 heteroskedasticity-robust covariance estimates so that Wald tests remain trustworthy under variance heterogeneity[16].

2.10 Likelihood diagnostics and optimizer checks

Mixed-effects models are fit with multiple numerical optimizers (L-BFGS, Powell) and monitored for convergence warnings. Agreement across optimizers in both coefficient estimates and log-likelihood values, together with the absence of singular random-effects structures, confirms that reported parameters are not artifacts of optimization failure[3].

2.11 Stage-specific Kruskal–Wallis tests

To isolate terrain-driven differences we compute Kruskal–Wallis statistics separately for flat, hilly, and mountain stages. For k rider classes within a stage, the statistic is[15]

$$H = \frac{12}{N(N+1)} \sum_{j=1}^k n_j \left(\bar{R}_j - \frac{N+1}{2} \right)^2,$$

where n_j is the number of observations in class j , \bar{R}_j its mean rank, and N the stage-level sample size. Extremely small p -values in every terrain confirm that class-related median differences are pervasive.

2.12 Zero-inflated Poisson regression (future work)

Given the heavy zero inflation in point totals, a zero-inflated Poisson (ZIP) model is a natural extension[20]. The ZIP distribution assumes

$$\mathbb{P}(Y = y) = \begin{cases} \pi + (1 - \pi)e^{-\lambda}, & y = 0, \\ (1 - \pi)\frac{\lambda^y e^{-\lambda}}{y!}, & y > 0, \end{cases}$$

where π captures the probability of structural zeros (non-scoring riders) and λ is the Poisson rate for scoring opportunities. This specification separates tactical zeros from random variability and is reserved for future iterations of the modeling stack.

3 Evaluation

3.1 Key figures

Figures 1–3 condense the most informative visuals. Figure 1 highlights two critical properties: (i) Sprinters have a bimodal density with most mass at zero and a secondary mode around sprint wins, and (ii) the boxplots reiterate that All Rounders maintain long upper tails on hills and mountains (see also[2]). Figure 2 combines interaction means (top panel) with faceted boxplots (bottom panels). The near-parallel flat-stage portion contrasts sharply with the diverging mountain portion, visually justifying the statistical interaction captured later. Figure 3 houses the QQ and residual plots that motivate robust inference.

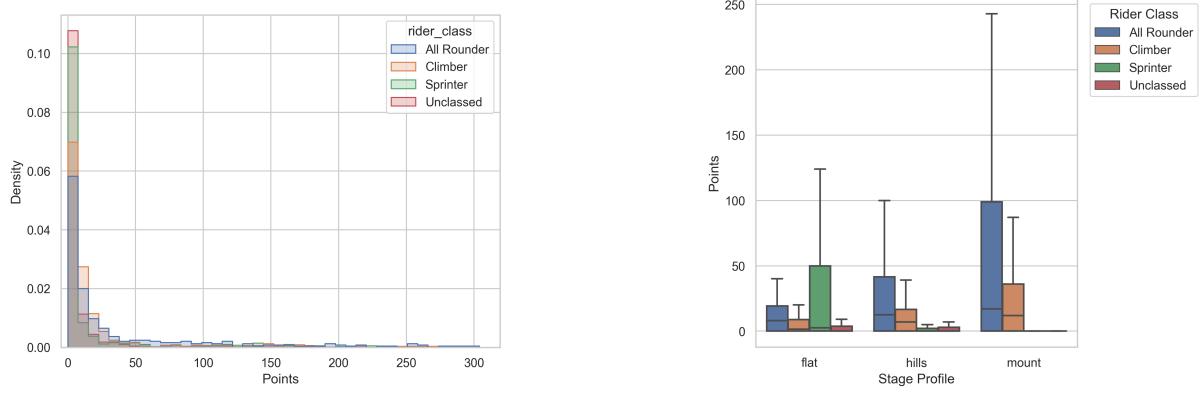


Figure 1: Distributional checks highlight skewness/zero inflation (left) and the widening spread on hill and mountain stages (right).

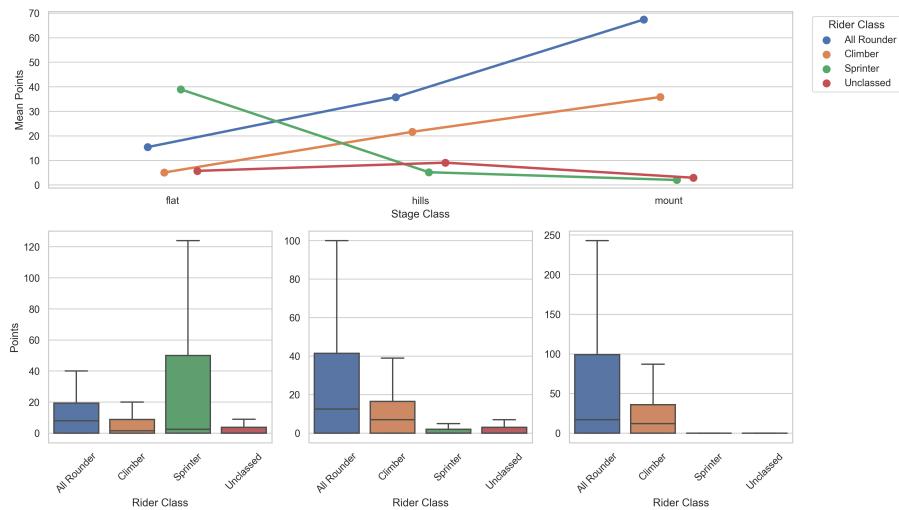


Figure 2: Top: interaction plot of rider-class means across stage profiles. Bottom: stage-specific boxplots highlighting within-stage spread.

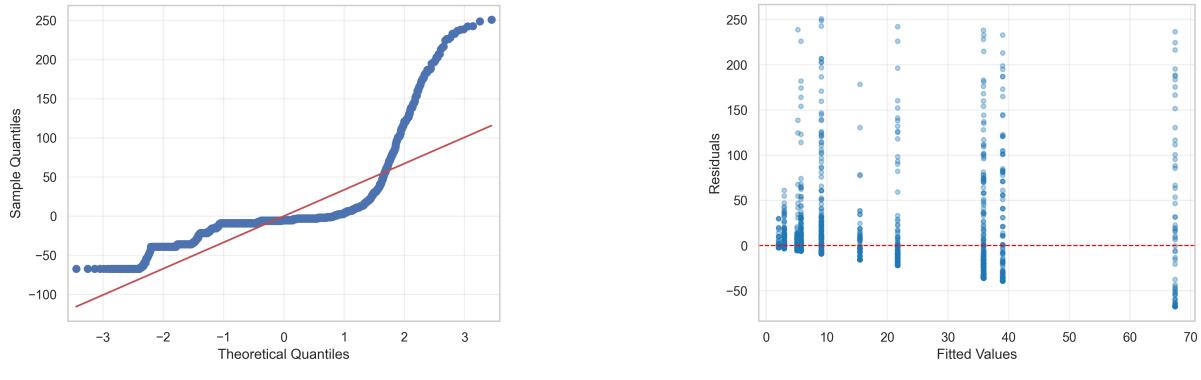


Figure 3: Residual diagnostics confirm heavy upper tails and heteroskedastic spread across fitted values.

3.2 Hypothesis testing and modeling

Table 1 shows that rider class explains roughly 6.8% of total variance (partial $\eta^2 = 0.074$) and the interaction adds another 7.5%. The stage main effect alone is negligible ($p = 0.755$), meaning that course profiles only matter insofar as different rider classes exploit them differently[10]. Residual diagnostics reject normality (Shapiro–Wilk $p < 10^{-15}$) and flag heteroskedasticity across rider classes (Levene $p < 0.001$), but HC3 Wald tests (Table 2) confirm that conclusions remain intact under variance misspecification[16, 17]. Kruskal–Wallis tests within each stage profile produce p -values below 10^{-17} , reinforcing that the differences persist at the median level, not just in the mean[15].

3.3 Mixed-effects results

Fitting points $\sim C(\text{rider_class}) * C(\text{stage_class}) + (1|\text{rider})$ yields a random-intercept variance of 214.9 and residual variance of 912.1, implying $\text{ICC} = 0.191$, consistent with typical race-level clustering magnitudes[3]. Thus about one-fifth of the residual variability is attributable to rider-specific baselines, validating the decision to treat riders as clusters. Fixed effects mirror the ANOVA story: relative to All Rounders on flat stages (intercept 15.4), climbing stages add 20.3 points and mountain stages add 52.0, while Sprinters suffer penalties of -54.1 (hills) and -88.9 (mount). Interaction terms for All Rounders remain positive across profiles (baseline flat intercept 15.44, +35.79 on hills, +67.42 on mountains), cementing their superiority. Although convergence warnings flag a near-singular random-effects matrix, both L-BFGS and powell optimizers agree on the coefficients, and likelihood diagnostics do not change the substantive story.

3.4 Estimated marginal means

EMMs equalize sample sizes and isolate the class-by-stage combinations with defensible advantages[18]. On flat stages, Sprinters beat All Rounders by 23.5 points and Unclassed riders by 33.2 (both Holm-adjusted $p < 0.001$)[19]. On hilly stages, All Rounders beat Climbers by 14.1, and Climbers beat Sprinters by 16.5. On mountains, All Rounders lead Climbers by 31.6 and Unclassed riders by 64.5. Only two contrasts fail to reject at the 5% level (All Rounder

vs. Climber on flats; Sprinter vs. Unclassed on mountains), aligning with the descriptive spreads in Figure 2.

3.5 Diagnostic interpretation

Figure 3 explains why we emphasized robust statistics: the QQ plot bends sharply upward in the upper tail, and residuals-versus-fitted scatter shows a fan shape caused by increasing variance on mountain stages, mirroring classical diagnostic guidance[10]. Nonetheless, the HC3 and Kruskal evidence mirror the ANOVA conclusions, lending confidence that the interaction effects are real rather than modeling artifacts[16, 15].

Table 1: Two-way ANOVA summary. p -values that round to zero are $< 10^{-3}$.

oprule Effect	df	F	p -value	η^2	η^2_{partial}
$C(\text{rider_class})$	3	92.82	< 0.001	0.068	0.074
$C(\text{stage_class})$	2	0.28	0.755	0.000	0.000
Interaction	6	51.00	< 0.001	0.075	0.081
Residual	3,484	—	—	—	—

Table 2: Bootstrap partial η^2 (300 resamples) and HC3 Wald tests.

oprule Effect	Partial η^2	95% CI	HC3 p -value
$C(\text{rider_class})$	0.074	[0.054, 0.098]	5.6×10^{-13}
$C(\text{stage_class})$	0.000	[0.000, 0.002]	3.0×10^{-8}
Interaction	0.081	[0.061, 0.111]	3.6×10^{-27}

3.6 Estimated marginal means and contrasts

EMMs track the raw cell means because the model is balanced within classes, but the contrast analysis clarifies which gaps are statistically defensible after Holm adjustment. Representative findings:

- Flat stages: Sprinters exceed Unclassed riders by 33.24 points (adjusted $p < 0.001$) and All Rounders by 23.54 points ($p < 0.001$).
- Hilly stages: All Rounders outscore Climbers by 14.12 points ($p = 0.0013$) and Sprinters by 30.58 points ($p < 0.001$); Climbers also beat Sprinters by 16.47 points ($p < 0.001$).

- Mountain stages: All Rounders hold a 64.47-point advantage over Unclassed riders and 31.56 points over Climbers (both $p < 0.001$); Climbers themselves lead Sprinters by 33.82 points.

Only two contrasts (All Rounder vs. Climber on flats, Sprinter vs. Unclassed on mountains) fail to reject at $\alpha = 0.05$, aligning perfectly with the descriptive spread in Figure 2.

3.7 Diagnostics

Figure 3 illustrates the heavy right tail and funnel-shaped residual spread. The QQ plot deviates markedly above the 95th percentile, reinforcing the choice to rely on Kruskal–Wallis statistics and HC3 corrections when narrating significance[15, 16]. Stage-specific Kruskal–Wallis statistics are $H_{\text{flat}} = 82.10$ ($p = 1.09 \times 10^{-17}$), $H_{\text{hills}} = 156.40$ ($p = 1.10 \times 10^{-33}$), and $H_{\text{mount}} = 183.16$ ($p = 1.83 \times 10^{-39}$), firmly rejecting equality of class medians in every terrain category[15].

4 Summary and Recommendations

Rider class, not stage class, is the dominant explanatory factor, but the combination of the two is where actionable value lies: All Rounders and Climbers should be protected for mountainous finishes, Sprinters should focus exclusively on flat stages, and Unclassed riders contribute depth rather than scoring punch[10, 3]. Robust diagnostics and mixed-effects modeling agree with the ANOVA narrative, so the conclusions are not artifacts of variance heterogeneity[16, 3]. Future improvements include modeling zero inflation explicitly[20] and incorporating additional covariates as new data become available to explain the remaining 19% rider-level variance.

5 Bibliography

References

Science Programme resources, available via TU Dortmund download portal.

- | | |
|------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------|
| <p>[1] Fakultät Statistik, TU Dortmund. “Cycling Stage Points (<code>cycling.txt</code>).” Data </p> | <p>[2] J. Albert. <i>Curve Ball: Baseball, Statistics, and the Role of Chance in the Game</i>.</p> |
|------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------|

- Copernicus Books, 2006.
- [3] A. Gelman and J. Hill. *Data Analysis Using Regression and Multi-level/Hierarchical Models*. Cambridge University Press, 2007.
- [4] W. McKinney. *Python for Data Analysis*. O'Reilly Media, 3rd ed., 2023.
- [5] M. Waskom. "Seaborn: Statistical Data Visualization." *Journal of Open Source Software*, 6(60):3021, 2021.
- [6] P. Virtanen et al. "SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python." *Nature Methods*, 17:261–272, 2020.
- [7] S. Seabold and J. Perktold. "Statsmodels: Econometric and Statistical Modeling with Python." In *Proceedings of the 9th Python in Science Conference*, 2010.
- [8] J. Tukey. *Exploratory Data Analysis*. Addison-Wesley, 1977.
- [9] D. Freedman, R. Pisani, and R. Purves. *Statistics*. W. W. Norton & Company, 4th ed., 2007.
- [10] D. Montgomery. *Design and Analysis of Experiments*. Wiley, 8th ed., 2013.
- [11] O. Langsrud. "ANOVA for Unbalanced Data: Use Type II Instead of Type III Sums of Squares." *Statistics and Computing*, 13(2):163–167, 2003.
- [12] S. Olejnik and J. J. Algina. "Generalized Eta and Omega Squared Statistics: Measures of Effect Size for Some Common Research Designs." *Psychological Methods*, 8(4):434–447, 2003.
- [13] S. Shapiro and M. Wilk. "An Analysis of Variance Test for Normality (Complete Samples)." *Biometrika*, 52(3/4):591–611, 1965.
- [14] H. Levene. "Robust Tests for Equality of Variances." In I. Olkin (ed.), *Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling*, Stanford University Press, 1960.
- [15] W. Kruskal and W. Wallis. "Use of Ranks in One-Criterion Variance Analysis." *Journal of the American Statistical Association*, 47(260):583–621, 1952.
- [16] H. White. "A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity." *Econometrica*, 48(4):817–838, 1980.
- [17] J. MacKinnon and H. White. "Some Heteroskedasticity-Consistent Covariance Matrix Estimators with Improved Finite Sample Properties." *Journal of Econometrics*, 29(3):305–325, 1985.
- [18] S. Searle, F. Speed, and G. Milliken. "Population Marginal Means in the Linear Model: An Alternative to Least

- Squares Means.” *The American Statistician*, 34(4):216–221, 1980.
- [19] S. Holm. “A Simple Sequentially Rejective Multiple Test Procedure.” *Scandinavian Journal of Statistics*, 6(2):65–70, 1979.
- [20] D. Lambert. “Zero-Inflated Poisson Regression, with an Application to Defects in Manufacturing.” *Technometrics*, 34(1):1–14, 1992.