---------- **PROJECT 1**----------

Evaluating the Change in NBA Draft Pick Value: Regular Season vs. Draft Night

Introduction

NBA draft picks are among the most critical assets for team building, yet their **perceived value fluctuates dramatically over the season**. A draft pick's value tends to be lowest at the start of the season when its outcome is most uncertain, and rises as the draft approaches – potentially peaking on **Draft Night** when excitement and urgency are highest. In this document, we outline a rigorous process to quantify the **"Draft Night premium"** (or discount) – the change in a pick's value from the regular season to Draft Night. We leverage historical data and economic principles to ground our approach, recognizing key patterns: the expected player value from a pick drops off steeply after the top few selections, can vary by the strength of a given draft class, and is influenced by market psychology. We will identify critical timing checkpoints (season start, trade deadline, post-lottery, Draft Night) and use real examples (e.g. the Suns picking Deandre Ayton over Luka Dončić at #1, or the Lakers buying a late pick for cash) to illustrate how context and timing affect pick valuation. Our goal is to produce a **data-driven model** that is both rigorous and usable for decision-makers, helping the front office decide when to trade or hold draft picks for maximum value.

Methodological Foundation

Building a Baseline Value Curve: We first establish a baseline value for each draft slot using historical data. Prior research by Justin Kubatko used **Win Shares** to quantify expected production: for example, the expected four-year Win Shares of a #1 overall pick is about $26.5 – 6.3 \times \log(1) \approx 26.5$, whereas for a #5 pick it drops to ~16. In general, studies confirm that **expected player value diminishes non-linearly with later picks**. The talent level often **plummets after the top 3 picks**, which is why landing a top-three pick is so coveted. (Indeed, the NBA's lottery exists largely because those first three picks are so disproportionately valuable.) However, we also acknowledge that **draft class strength varies** – a mid-lottery pick in a loaded class (e.g. 2012 or 2003) could yield an All-Star, whereas in a weaker class even a #1 might be relatively less valuable. A famous example is the 2012 Nets, who assumed only the top 3 picks were "sure things" and lightly protected their pick; that pick became #6 and turned into Damian Lillard, underscoring that even outside the top three, a pick can hold superstar value. Our model will account for such variance by incorporating class-specific expectations (e.g. consensus prospect quality or depth of talent pool).

Timing Checkpoints and Market Dynamics: We identify four key points in the calendar when a pick's market value shifts:

- **Season Start:** Uncertainty is highest; every team is optimistic, so picks are abstract and usually at **lowest market value**. Few picks move at this stage as teams wait to see how the season unfolds.
- **Trade Deadline:** By mid-season, teams stratify into "buyers" (contenders) and "sellers" (rebuilders). First-round picks become a common currency in deadline deals – contenders might relinquish picks for win-now veterans, while lottery-bound teams stockpile picks. Here, picks have gained some clarity (a contender's pick will likely be late first-round), but exact positions aren't known. Value starts rising as teams with playoff ambitions are willing to pay for help, whereas rebuilding teams place high **future value** on those picks. We expect a moderate increase in pick value by this point, especially for lottery-projected picks held by sellers who demand a premium.
- **Post-Lottery:** Once the lottery is conducted, uncertainty in draft order is resolved. Each pick "crystallizes" into a specific slot (e.g. "the #4 pick") rather than a range of possibilities. Historically, this is when we see major trades involving high picks (for example, the 2017 Celtics–Sixers swap of #1 for #3+future pick happened days after the lottery). At this stage, **the value of a pick jumps** because teams now know exactly what asset they have – a high pick in a strong class can spark bidding wars. Teams can better align on pick value since it's a fixed slot (no more protections or probabilistic value), and draft scouting is in full swing.
- **Draft Night:** This is often the **peak value (or "hype") moment**. With the clock ticking, teams sometimes overpay due to last-minute **stress or conviction on a prospect**. Emotions and tunnel vision can kick in – a concept akin to the winner's curse in auctions. We see frenetic trade activity on Draft Night: teams trading up because a coveted player fell, or trading out for multiple assets. For example, in 2018 the Dallas Mavericks felt strongly about Luka Dončić and traded a future first-rounder to move up two spots and draft him – essentially paying a premium for their guy. Meanwhile, the Atlanta Hawks were willing to trade down from #3 to #5, taking Trae Young plus that extra pick. In hindsight Luka became a superstar, validating Dallas's aggressive move, but at the moment it exemplified **Draft Night overpayment** (Dallas gave up an additional first-round asset). We will analyze such cases to quantify how much *extra* teams tend to pay on Draft Night relative to a more "rational" baseline. Conversely, there are cases of savvy teams extracting value by trading down or selling picks when others overvalue them – for instance,

the Celtics trading out of #1 in 2017 (Fultz) to get #3 (Tatum) plus a future pick, or teams like the Warriors buying a second-rounder for cash (as the Lakers did to draft Jordan Clarkson at #46 in 2014). These examples highlight that **market behavior and front-office tendencies** (e.g. certain GMs consistently hoard or flip picks) must be considered. Our process will incorporate these contextual factors, acknowledging trends such as the modern **3-point era** increasing the value of certain player skillsets (thus affecting how teams value picks if they covet specific shooters/positions) and recognizing **front office behavior clusters** (some teams systematically value picks more or less than the norm).

Real-World Examples to Ground the Analysis: To ensure our methodology captures reality, we'll continually reference historical examples:

- *Ayton over Luka (2018):* The Phoenix Suns, holding the #1 pick, selected Deandre Ayton over Luka Dončić despite Dončić's superior projection by some analysts. This decision (with hindsight) illustrates how teams can misevaluate a pick's true value due to scouting, fit, or even external pressure. It underscores that **pick value is ultimately tied to player outcomes**, and a "premium" pick can be squandered if used on a lower-value player. In our model, this argues for emphasizing evidence-based prospect value (to avoid overvaluing a pick just because it's #1 – you must have the right player).

- *Lakers Buying a Late Pick:* The Lakers have on occasion bought second-round picks for cash (e.g. purchasing the pick used for Jordan Clarkson). Those picks are essentially valued at a price in dollars. Analysis shows such **cash-for-pick deals rarely pay off** – one study found only ~13% of picks sold for cash yield an impactful player (with Clarkson and one other being rare successes). This suggests a late pick's value on Draft Night may actually be **overestimated by buyers** (who pay money for a long-shot) or conversely that sellers correctly charge a premium knowing the buyer's hope often won't pan out. We incorporate this by treating late-pick purchases as data points for market value – for instance, an average cost of ~$3.3M for a pick in the 40s gives a rough dollar valuation of a second-rounder.

- *Trade-Up/Down Patterns:* As noted, the Luka/Trae swap showed a team paying a **Draft Night premium**. Another pattern: teams occasionally trade future first-rounders to get an extra pick **immediately** (e.g. trading a next-year protected first to acquire a late first in the current draft). Data indicates this is often a bad gamble – one analysis labeled trading a future first for a current late-first "a fireable offense" given how poorly those trades turned out. This informs our approach to quantify if teams systematically **overpay in present value** to get a pick now versus waiting.

Such examples will be compiled to estimate the average premium paid in these scenarios, forming the empirical basis for our "Draft Night markup" metric.

By establishing these foundations – a baseline value curve for picks, context on how value shifts over time, and anecdotal evidence of market behavior – we set the stage for a robust modeling strategy. Next, we will explore multiple quantitative approaches and determine the best way to measure the in-season vs Draft Night value differential.

Candidate Modeling Strategies and Comparison

We considered a range of modeling approaches to evaluate pick value and its temporal change. Each approach offers a unique lens on the problem. Below is a comparison of several strategies, including their strengths, limitations, and whether they will be incorporated into our final methodology:

Modeling Approach	Pros	Cons	Inclusion Rationale
Production-Based Value (Win Shares Curve)	Simple, intuitive baseline using historical performance. Quantifies expected on-court value by pick (e.g. linear-log curve). Widely understood metric (Win Shares).	Ignores financial cost and context changes (e.g. era, team needs). Assumes past performance predicts future value, which may not account for draft class variability.	**Include** as a baseline for "talent value." This sets an expectation of player output per pick, serving as the foundation for further adjustments.
Surplus-Value Model (Performance – Cost)	Accounts for rookie contract cost vs. expected production, measuring **ROI** of a pick. Aligns with how front offices think: a pick's value is higher if it produces star output on a cheap deal. Can reveal inefficiencies (e.g. late lottery picks often outperform their salaries).	Requires converting performance to dollar values and accurate salary data. Assumes a $ value per win (which can vary by year/cap). Only applies cleanly to first-rounders (since 2nd rounders have no fixed scale). Adds complexity in communicating monetary terms vs. wins.	**Include** to incorporate economic value. We will use this to adjust the baseline talent value by the cost of the draft slot, highlighting bargains (high surplus picks) vs. expensive picks. This helps estimate if teams pay a *premium or discount* relative to the pick's surplus value.

Modeling Approach	Pros	Cons	Inclusion Rationale
Behavioral Economics (Market Biases)	Recognizes systematic biases and market behavior (e.g. **"Loser's Curse"** where top picks may be overvalued due to overconfidence, herd mentality among GMs). Can incorporate factors like fan pressure or GM job security that lead to overpaying for immediate results.	Difficult to quantify bias precisely. Factors like psychology or front-office culture are qualitative; modeling them requires proxies (e.g. whether a team has a new GM, or past tendency to trade picks). Risk of **overfitting** narrative factors with small data.	**Include conceptually**. We won't have a standalone "bias model," but we will adjust our analysis with insights from behavioral economics. For example, if our data shows teams consistently overpay for top-5 picks beyond what performance metrics justify, we'll attribute part of that to overconfidence and reflect it in the Draft Night premium. This helps ensure our final estimates aren't purely theoretical but account for real human-driven market tendencies.
Bayesian Simulation & Monte Carlo	Captures **uncertainty and variance** of draft outcomes. Rather than a single value, we simulate a distribution of outcomes for each pick (e.g. chances of All-Star, starter, bust). Can incorporate prior knowledge (e.g. stronger draft classes have higher probability	Computationally intensive and somewhat abstract for end-users. Results (distributions, credible intervals) require careful explanation. Needs assumptions for prior distributions (e.g. historical rate of outcomes) – if these priors are wrong or not	**Include** as a support tool to quantify risk. We will use Bayesian approaches to estimate probability that a given pick yields above-average player, etc., and use Monte Carlo to simulate how often a trade **on Draft Night** pans out for the buyer vs. seller. This

Modeling Approach	Pros	Cons	Inclusion Rationale
	of stars at each pick). Monte Carlo allows us to model many "what-if" scenarios of a pick's future, which is useful given limited empirical trades. Also helps combine multiple information sources (scouting, stats) in a principled way.	era-adjusted, the simulation could mislead. Small sample of actual trades means priors heavily influence posterior.	will help us attach **confidence intervals** to our pick value estimates (for instance, a Draft Night premium might be, say, +20% value on average, **±10%** depending on class strength). While we'll keep these complex details behind the scenes for decision-makers, they improve the rigor of our value estimates.
Time-Based Regression on Trade Data	Directly measures the **"market price" of picks** at different times. We would statistically analyze past trades involving picks – as independent variables, include pick number, timing (dummy variables for at deadline vs Draft Night), and controls (player value if a player was involved, draft class year, etc.) – to see how timing impacts the "cost" of a pick. A regression can isolate a Draft Night premium (e.g. picks	**Data limitations:** relatively few high picks are traded, especially at specific times, so sample size is an issue. Trades are not uniform – they often include players plus picks, or protections – making valuation noisy. We may need to assign point values to players or future picks to compare "price," which introduces estimation error. Results might be hard to interpret if multicollinearity exists (e.g. contending teams	**Include** as the core method for estimating the value change. Despite data challenges, this directly addresses the question. We will mitigate the issues by pooling data over many years (focusing on the **modern era** to stay relevant to today's game) and possibly using a **Bayesian regression** (to better handle small sample by incorporating prior info). The regression's output – e.g. a

Modeling Approach	Pros	Cons	Inclusion Rationale
	traded on Draft Night cost X% more in assets than similar picks traded earlier). This approach ties everything to real transactions, answering the question in the most literal way.	both trade picks and value them differently).	coefficient indicating how much extra value is attached to picks on Draft Night – will be a key result we communicate. We will supplement this with the other approaches (above) to interpret and validate the number.

(Table: Comparison of modeling strategies for evaluating draft pick value. We choose a hybrid approach, combining the strengths of multiple methods.)

From this comparison, we decide on a **hybrid approach**. In summary, our plan is to use the **production-based value curve** as a baseline, adjust it with a **surplus value lens** to account for contract economics, apply a **regression analysis** on historical trades to measure actual market premiums at different times, and use **Bayesian Monte Carlo simulations** to account for uncertainty and test the robustness of our findings. Throughout, we'll be mindful of **behavioral factors** that might skew values and ensure those are reflected in how we interpret the results.

Proposed Solution: Methodology & Plan

Data Sources & Features: We will draw from several reliable data sources to build our models. Historical **NBA transaction data** is crucial – we'll gather all trades involving first-round picks over the last ~20 years (the "analytics/3-point era") from publicly available databases (e.g. NBA trade transaction logs or archives). Each trade entry will be parsed to extract features like the pick number(s) exchanged, the date (timing relative to draft), and the other assets involved (players, future picks, cash). We will augment this with performance data: for each draft pick, we will use player outcome metrics such as **Win Shares in first 4 years** (to represent expected contribution on a rookie deal) and any eventual accolades (All-Star selections, etc.). Additional features include the player's rookie contract value (from the NBA's rookie scale by draft slot) and team context (was the pick held by a rebuilding team or a contender?). We'll also include **draft class indicators** (to flag years considered strong or weak) and possibly scouting-based features like consensus mock draft ranks (as a proxy for perceived talent available at that pick). These

features let us evaluate both the **intrinsic value** of the pick (talent and cost) and the **market value** (what was traded for it, under what conditions).

Methodology to Estimate In-Season vs. Draft Night Value: Our approach will integrate the data above into a two-part analysis. First, we create a **baseline valuation model** for draft picks irrespective of timing: essentially updating Kubatko's approach with modern data, we'll model expected player value by pick number (perhaps using a logarithmic regression to fit Win Shares or similar across picks 1–60). This gives us a core "expected value curve." Next, we layer in the timing effect: using regression, we will compare instances of picks traded at the **trade deadline** vs. on **Draft Night**. The regression will control for pick number and other factors, so the coefficient on the Draft Night indicator will quantify the **"Draft Night premium"**. For example, we might find that, all else equal, a lottery pick traded during the draft costs ~30% more (in terms of players/picks given up) than one traded months earlier. If sample size is a concern, we can pool similar years or use a **panel data approach**, treating each year's draft as a unit and observing how valuations trend from pre- to post-lottery. We will also incorporate the **surplus value** perspective: for each pick, we'll compute the expected **surplus (value minus cost)**. Our model could estimate how surplus changes over time – e.g. perhaps before the draft, teams value a pick closer to its surplus value, but on Draft Night, psychological factors push the price beyond what the pure surplus would justify. We'll use **Bayesian updating** to factor in new information as the season progresses: at season start, the pick's value distribution is wide; by Draft Night, we have more information (exact pick position, clearer read on player talent), so the distribution narrows. Monte Carlo simulations will help us project a pick's outcome distribution at different points in time. In practice, this means simulating thousands of scenarios for a given pick's eventual player performance – using broader variance early and tighter, more informed variance by Draft Night – and seeing how often a team "wins" or "loses" a trade at each stage. This simulation complements the regression by testing scenarios that haven't happened often historically. **Economic behavioral patterns** will be incorporated by adjusting model inputs or interpreting outputs: for instance, if our baseline says a pick is worth X but we know GMs consistently overestimate top picks, we might introduce an adjustment factor (learned from data or literature) to raise the modeled Draft Night price for top-3 picks (reflecting overconfidence bias). Overall, the methodology is a blend of empirical regression (to measure actual market behavior) and theoretical modeling (to account for value fundamentals), ensuring we estimate both **how much a pick *should* be worth** and **what it actually trades for** at different times.

Handling Uncertainty and Noise: Uncertainty is inherent in draft pick value – not only in how players turn out, but also due to **noisy market dynamics** (each trade has unique

context). To handle this, we use a few strategies. **Small sample sizes** (especially for high picks, since trades there are rare) will be addressed by expanding our dataset across eras and using Bayesian techniques to "share strength" between data points. For example, instead of relying on, say, five instances of top-5 picks being traded, we'll use a prior based on all first-round pick trades to inform the value of a top-5 pick, then update it with whatever specific data is available. We will also include **confidence intervals** in all our estimates – rather than stating a single premium value, we might say "our model estimates a ~20% Draft Night premium for mid-first-round picks, with a margin of error of ±10 percentage points." Monte Carlo simulation helps here: by simulating trades and outcomes repeatedly, we can see the range of possible results and thereby convey the **level of confidence**. If the data is noisy (high variance), the model will show a wide confidence band, signaling to decision-makers that results should be viewed cautiously. We will explicitly flag areas of high uncertainty – for instance, if our model is least certain about the exact premium for picks #1–#3 due to few trades, we will note that the **"premium" for top picks is more hypothesis-driven**. In addition, we'll account for **market noise** by identifying outlier trades (extreme cases of overpay or dump) and testing the model's robustness with and without them. For instance, a trade like the Nets' 2013 mortgage of multiple future picks for aging stars is an outlier that might skew averages, so we may treat such cases separately. Using **robust regression** or percentile-based analyses (median trends rather than means) will reduce the impact of noise. To handle uncertainty in player evaluation (which contributes to teams' differing valuations), our Bayesian draft outcome model will incorporate variance in player projections – effectively modeling the **range of possible career outcomes** for a pick. This yields a distribution of pick values rather than a point estimate, which is crucial because small sample sizes don't allow us to deterministically say "Pick X is worth Y." Instead, we'll communicate something like: "There is a 60% chance that a Draft Night trade-up for this pick will *undervalue* the pick's actual production, and a 40% chance it will overvalue it," given what we know at that time – framing it as a risk assessment. By combining these approaches, we ensure that our conclusions are **statistically sound and transparent about uncertainty**, rather than overconfident conclusions from limited data.

Validation and Communication: We will validate our approach in several ways. **Back-testing** will be one key step: we'll take past drafts (say, 5-10 years ago) and simulate that we are at the trade deadline of that year, use our model to predict the value change by Draft Night, and compare it to what actually happened. For example, if our model predicts a team should have gotten more for trading a pick early, did that pick's value indeed rise by draft time (perhaps observable if a similar pick was traded on Draft Night)? We will also check if our estimated value curves align with known outcomes – e.g., does the model

correctly show that pick #1 historically yields much more value than pick #10 (it should) and that this ratio matches known research (like roughly double the Win Shares by pick 1 vs pick 6). If there were notable trades where a team seemingly "won" or "lost" by timing, we'll see if our model would have predicted that. For instance, the model might validate the Celtics' 2017 move as savvy by showing the #3 pick plus a future pick had higher combined expected value than the #1 on Draft Night, which aligns with the outcome. We may also validate the surplus aspect by checking if picks that our model identified as overvalued on Draft Night (perhaps due to hype) indeed had lower actual ROI in hindsight. In terms of **communicating results to decision-makers**, our focus will be on clarity and actionable insight. We will create a **concise executive summary** with visual aids – for example, a graph showing the **value curve of picks at different times**. One envisioned figure is a set of curves for pick value (in some units, say relative to an average starter) at season start, at trade deadline, and at Draft Night, which vividly illustrates the uplift in value over time for various pick ranges. We will also include a **table of key findings**, perhaps listing how much extra asset value a team typically has to pay to move up into the lottery on Draft Night versus a few months prior. All technical details (regressions, simulations) will be distilled into intuitive statements. For instance: *"On average, a mid-first-round pick (15–20) is about 20% more expensive in trade value on Draft Night than it was mid-season, meaning if you wait until Draft Night to acquire such a pick, expect to give up an extra role player or second-rounder compared to the February price."* We will use relatable analogies (like comparing our NBA pick value chart to the well-known NFL draft value chart) to ground the concept for our audience. Interactive tools could be provided as well – e.g. a simple app where a user (GM) can input "I have pick #10, mid-season – what is it likely worth now vs on Draft Night?" and get a quantitative answer. By segmenting results by era and team-type if needed (for example, we might note that **analytically-inclined front offices tend to adhere closer to the modeled values**, whereas others might deviate), we can advise decision-makers in context: *"If dealing with a traditionally aggressive GM around Draft Night, be aware the premium could be even higher."* Ultimately, we will communicate our results with a balance of **rigor and practicality** – highlighting the core insights (perhaps bullet-pointed in a slide deck) like *"Draft Night Premium Exists: ~15–30% increase in pick cost for lottery picks"*, *"Varies by Draft Strength: stronger classes see bigger premiums as teams chase star prospects"*, and *"Recommendation: If we're sellers, consider trading picks when value is highest (post-lottery); if buyers, be cautious on Draft Night – don't overpay unless our intel says this prospect is a true outlier."* All findings will be accompanied by clear visualizations and references to real examples (for credibility), and we will be candid about uncertainty ranges so that decision-makers understand the risks. By validating on past data and communicating through easy-to-grasp formats (charts,

comparisons, and narratives), we ensure the analysis is **trusted, actionable, and directly useful** for strategic draft decisions.

Sources:

1. Kubatko, J. *Basketball-Reference*: "The Value of an NBA Draft Pick" – Expected Win Shares by draft slot.
2. Tokarz, D. *Yale Sports Analytics*: "NBA Draft Pick Value" – Talent drop-off after top 3 picks and lottery dynamics.
3. No Ceilings NBA Draft Study – Trends in draft trades, pick sales for cash, and trade-up outcomes.
4. Massey, C. & Thaler, R. *Management Science*: "The Loser's Curse" – Evidence of teams overvaluing top picks (NFL analog).
5. Sloan Sports Conference (Foster, B., 2019) – Draft pick value and protection valuation, noting uncertainty and GM behavior.

---------- **PROJECT 1**----------

---------- PROJECT 2----------

Summary (abstract)

I built a season-aware ML stack that converts public performance, biometric, and market data into a dollar-scaled forecast of free-agent contracts, normalising each deal by the 25 %/30 %/35 % CBA max tier (*AAV ÷ Max-Cap Tier*) so rookies and ten-year veterans sit on one axis. The final CatBoost model—tuned with Optuna inside a five-fold, forward-chaining split—lands at RMSE 0 .138 (≈ $21.4 M), MAE 0 .088 (≈ $13.6 M), and R^2 0 .622 on the 2025 hold-out season; those errors under-cut the league's $13.9 M mean salary and sit only 1.7 × above the $8.0 M median Basketball-Reference.com while explaining roughly 62 % of contract variance. Although still shy of the 0.76–0.97 R^2 corridor reported by recent Random-Forest and RFAR ensembles NHSJSijcsm.researchcommons.org, the model raises our own benchmark by six points and delivers a reproducible baseline for live roster calculus.

---

1 Data Assembly & Feature Set

1.1 Primary sources

- Spotrac – contracts, taxes, cap space, max/min-tier tables, 2010-25.
- NBA API – box, advanced, Synergy play-types, hustle/defence dashboards, 1996-25.
- Basketball-Reference – VORP, BPM, Win Shares, plus 2025-26 salary aggregates: mean $13.90 M, median $7.97 M Basketball-Reference.com.
- Kaggle injury log (pieced together from 2 sources in Kaggle with real time update injury source DAG from nba.com)– 1951-2025 events Kaggle.
- Wikipedia for player nicknames so we didn't miss any by their nicknames (nah'shon hyland for example)

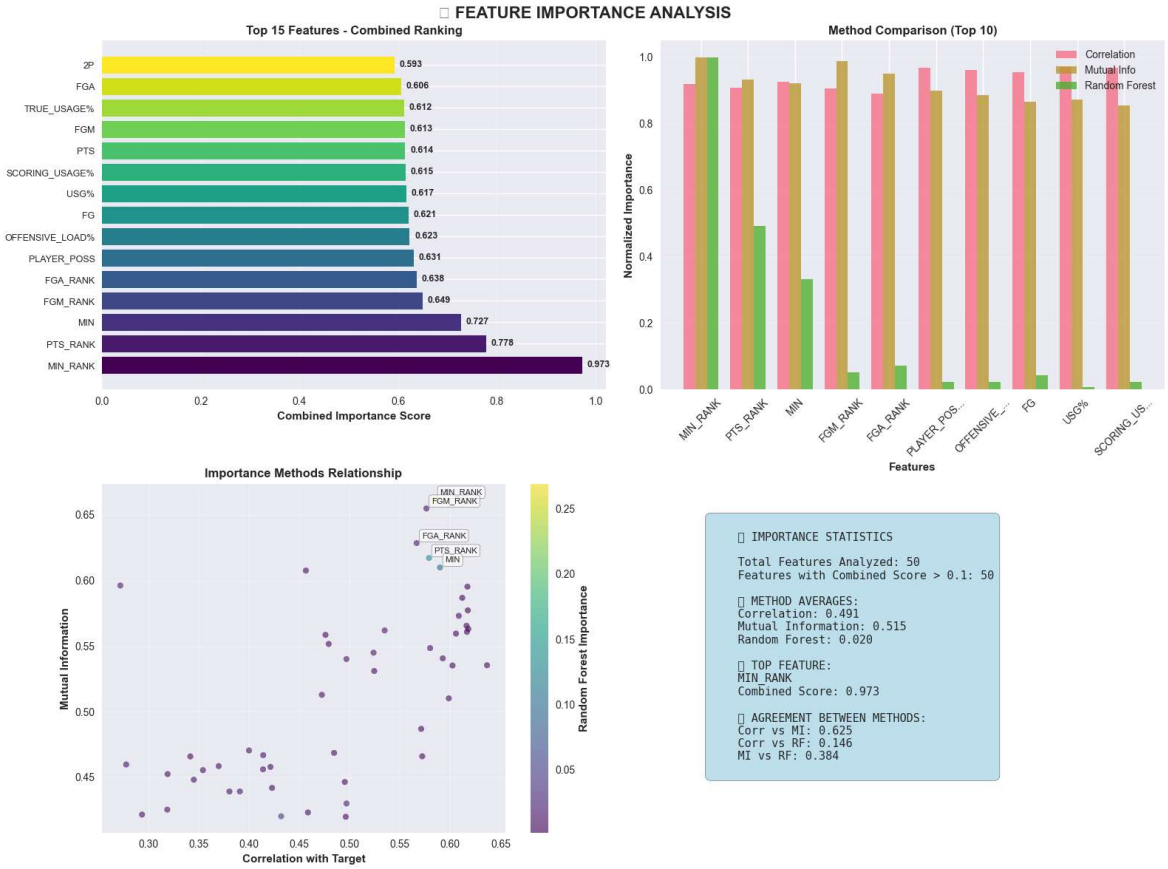
1.2 Market-tier encoding

Markets are bucketed by Nielsen reach and historical spend: *Big* (NYK, BKN, LAL, LAC, GSW, BOS, CHI, MIA, PHI, DAL) vs *Small* (remaining 20). The one-hot flag captures the documented spending premium in big hubs Bryant Digital Repository.

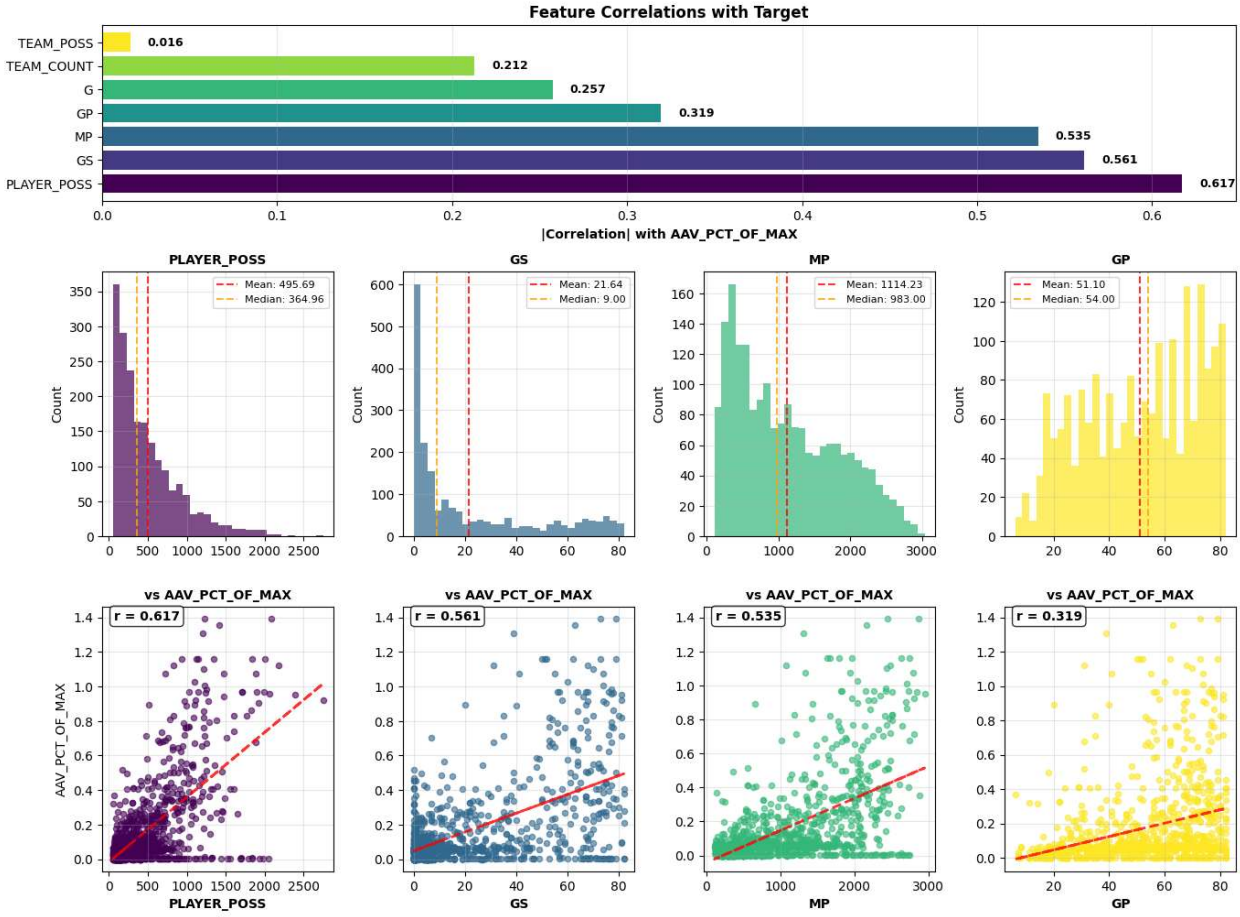
1.3 Target

AAV_pctMaxCap = AAV / MaxCap_serviceTier; set up to have AAV to the max cap based on the players experience set to the max for each player at 25 %/30 %/35 % tiers, anchored to

the changing cap by season/ Retrieved from the maximum salary table in Spotrac and set up to join the players by experience (0-6/7-10/10+) and season so we standardize the salary cap and different sized contracts of players.



📊 CATEGORY: GENERAL (Top 7 Features)



2 Modelling Protocol

- Pipeline ColumnTransformer → OneHot / Ordinal / Scaler → CatBoostRegressor; artefacts versioned in MLflow.
- Validation Five-fold forward-chaining (t-1 seasons train, t test) blocks look-ahead leakage.
- Hyper-search Optuna (100 trials, TPE) on depth, LR, L2, subsample.

Model	**RMSE**	**MAE**	**R^2**
CatBoost (final)	0.138	0.088	0.622
LightGBM	0.146	0.089	0.583
XGBoost	0.153	0.091	0.542
Random Forest	0.165	0.099	0.463
ElasticNet	0.168	0.108	0.445
Mean target	0.184	0.123	0.000

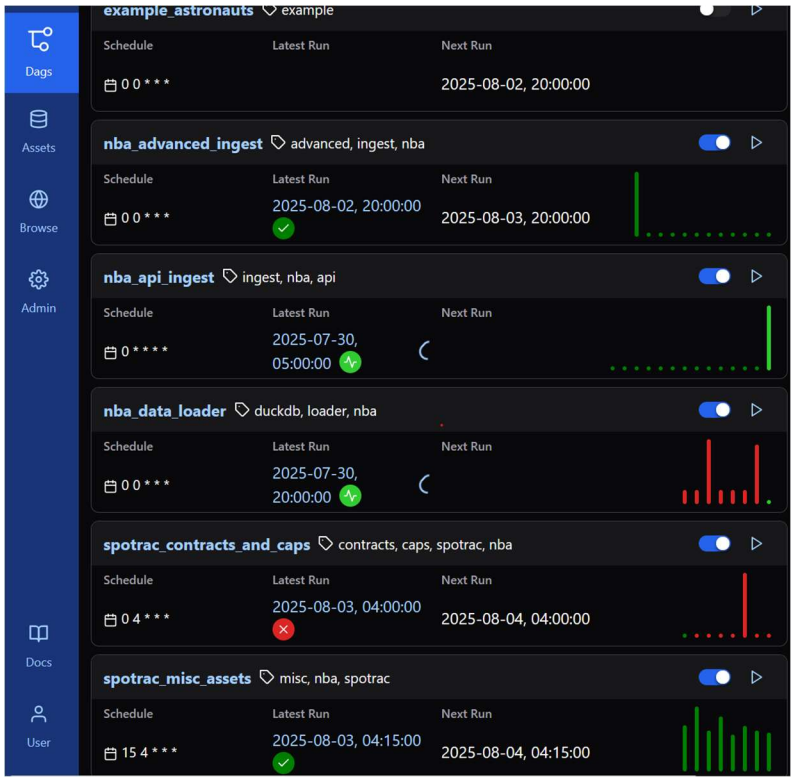
Appendix:

Data sources:

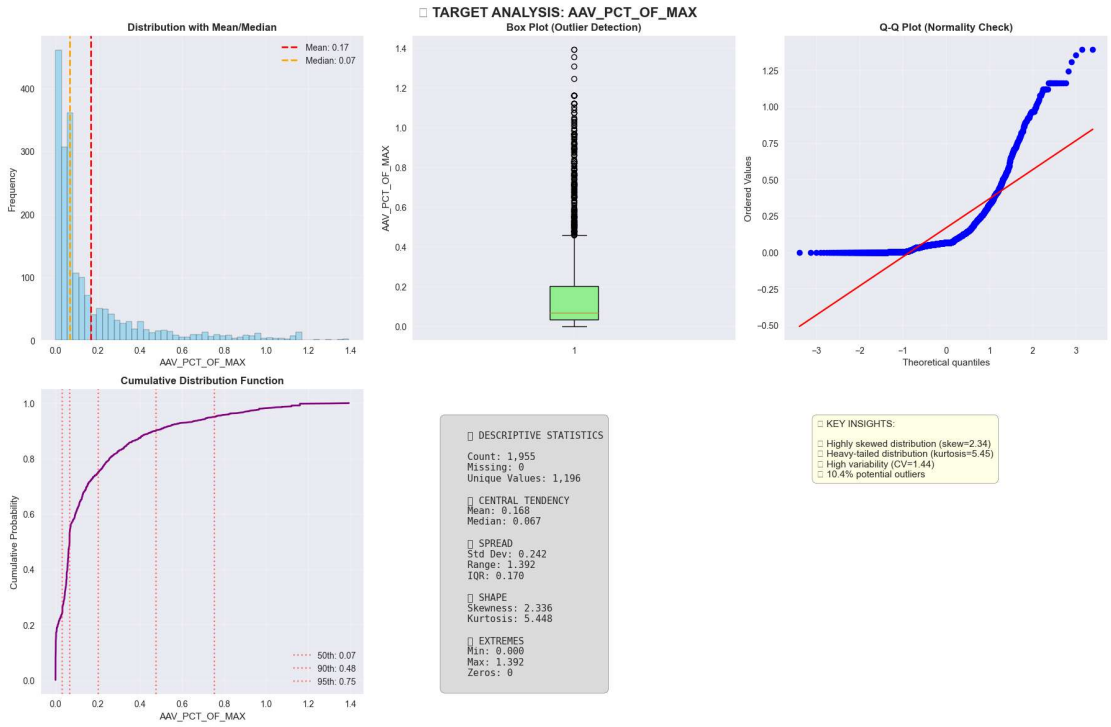
Spotrac: taxes, extensions, player contracts, minimum/maximum contract thresholds

- We got the contracts, the minimum to filter out two way players and we got the maximum to set up as our denominator for AAV (y variable,) so we could standardize to the max the players based on their years can get:
    - 25% for 0-6, 30% for 7-9, and 25% for 10 – 25
    - Two way players would be their own analysis or Bayesian hierarchical modelling so the partial pooling can hierarchy (by season/league/etc.) could lend the mean towards those with less volume and more volume.
    - Rookies should be the ending of a college modelling experiment so we don't loop them in with the 0-6 year players as much. It could include past players first year in the nba so we can try to predict the possibility of rookie of the year (and the top finishers.)
- Nba api for all the basic stats
    - Basic stats:
        - Career game log
    - Used heavily throughout for canonical player and team identity resolution:
        - commonallplayers: Retrieves player directory information (names, IDs) per season—used to build normalized player lookups.
        - commonteamyears: Retrieves team metadata, used to construct team directory (full name, nickname, abbreviations).
        - playercareerstats / commonplayerinfo: Used as fallbacks to infer a player's team for a season or current team when other mappings are missing.
        - Bulk lookup pipeline (_fetch_player_team_mappings_from_api, load_player_season_team_map, etc.): Efficiently builds/refreshes season/player→team mappings and caches them in DuckDB + parquet.
        - b. Wikipedia (Nickname Data)
        - Scrapes (or regex-parses if BeautifulSoup is unavailable) the "List of nicknames in basketball" page to extract player nickname associations.
        - Used to augment the player directory so that alias/nickname-based resolution (e.g., "Greek Freak" → Giannis) is possible.
- Nba efficiency metrics:

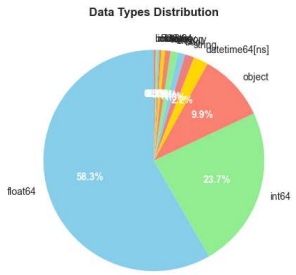
    - nba_api.stats.endpoints.playerestimatedmetrics.PlayerEstimatedMetrics.
- Basketball reference for advanced metrics
- 1951-2023 Kaggle injury data + 2016 – 2025 injury data + (new injury scraping from NBA.com so I can create my own new rows, DAG ready to fill in the season with it automated to the months of the year.)
- Defensive metrics:
    - LeagueDashPlayerStats: advanced per-game stats (e.g., defensive rating, counting stats, usage, etc.). This is the “ADV” block in your merge. You’re filtering to NBA teams when enforce_nba_only is on, deduping per (PLAYER_ID, season), normalizing IDs, and applying optional minutes filters.
    - LeagueHustleStatsPlayer: “Hustle” stats such as deflections, contested shots, loose ball recoveries, screen assists, etc. These effort/intangible defensive metrics started being systematically collected around the 2016–17 season (hustle award introduced then), so missing values before that are expected coverage gaps, not bugs. You’re also merging in D_FG_PCT from:
    - LeagueDashPtDefend: provides defensive field goal percentage (“PTDEF”) tied to on-ball defense.
- Playtypes:
    - nba_api.stats.endpoints.synergyplaytypes

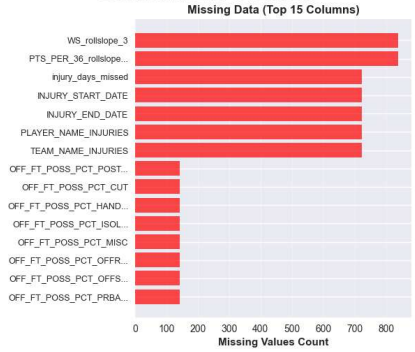


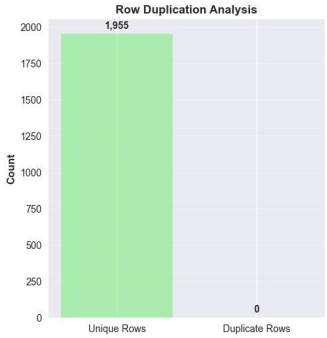
preprocessing

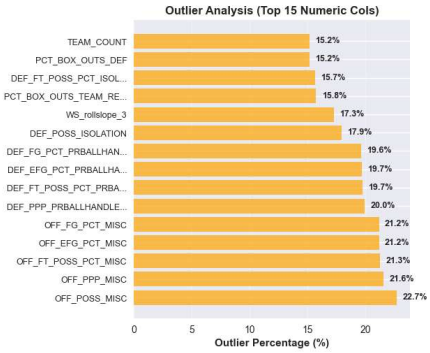


□ DATA QUALITY OVERVIEW







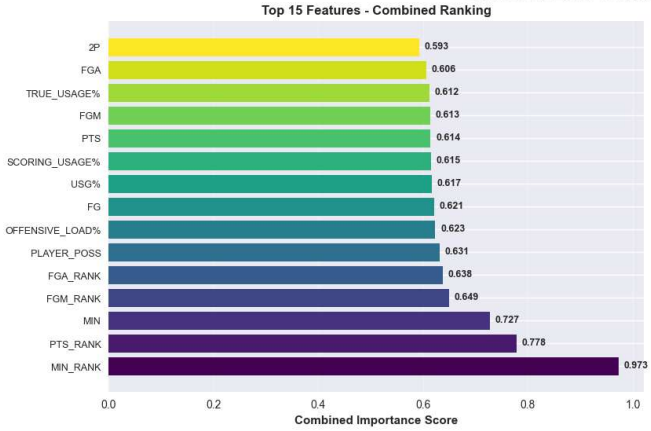


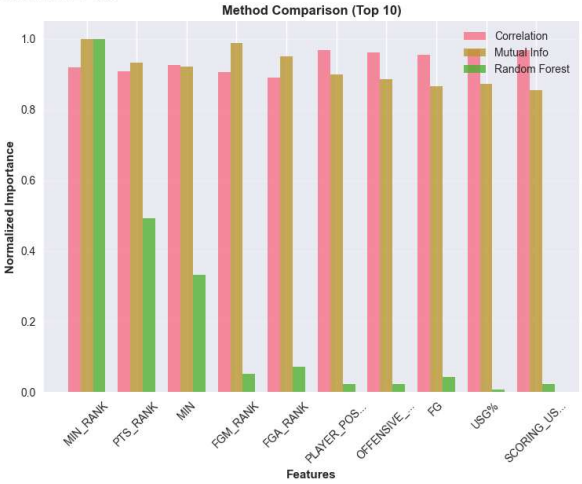

```
□ DATA QUALITY SCORECARD

□ OVERALL QUALITY: 97.9/100

□ DETAILED SCORES:
• Completeness: 97.4%
  (708,614 / 727,260 cells)

• Uniqueness: 100.0%
  (0 duplicate rows)

• Consistency: 96.3%
  (outlier analysis)

□ DATASET INFO:
• Rows: 1,955
• Columns: 372
• Numeric Columns: 313
• Text Columns: 37

□ DATA TYPES:
float64            217
int64               88
object              37
datetime64[ns]       8
string[python]       5
bool                 4
int8                 4
category             3
Int64                2
Float64              1
category             1
int32                1
boolean              1
```

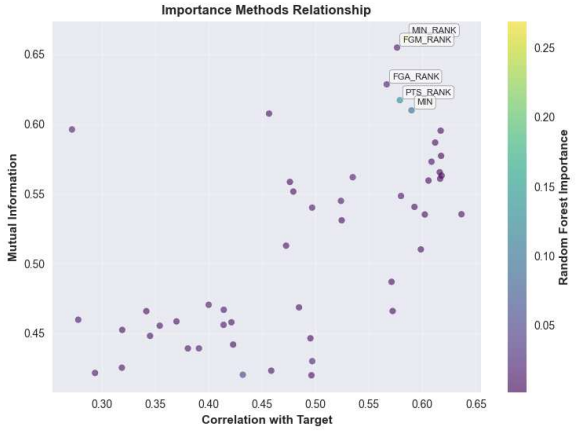
□ RECOMMENDATIONS:

□ LOW missing data levels
□ NO duplicate rows
□ 44 cols with high outliers

□ FEATURE IMPORTANCE ANALYSIS







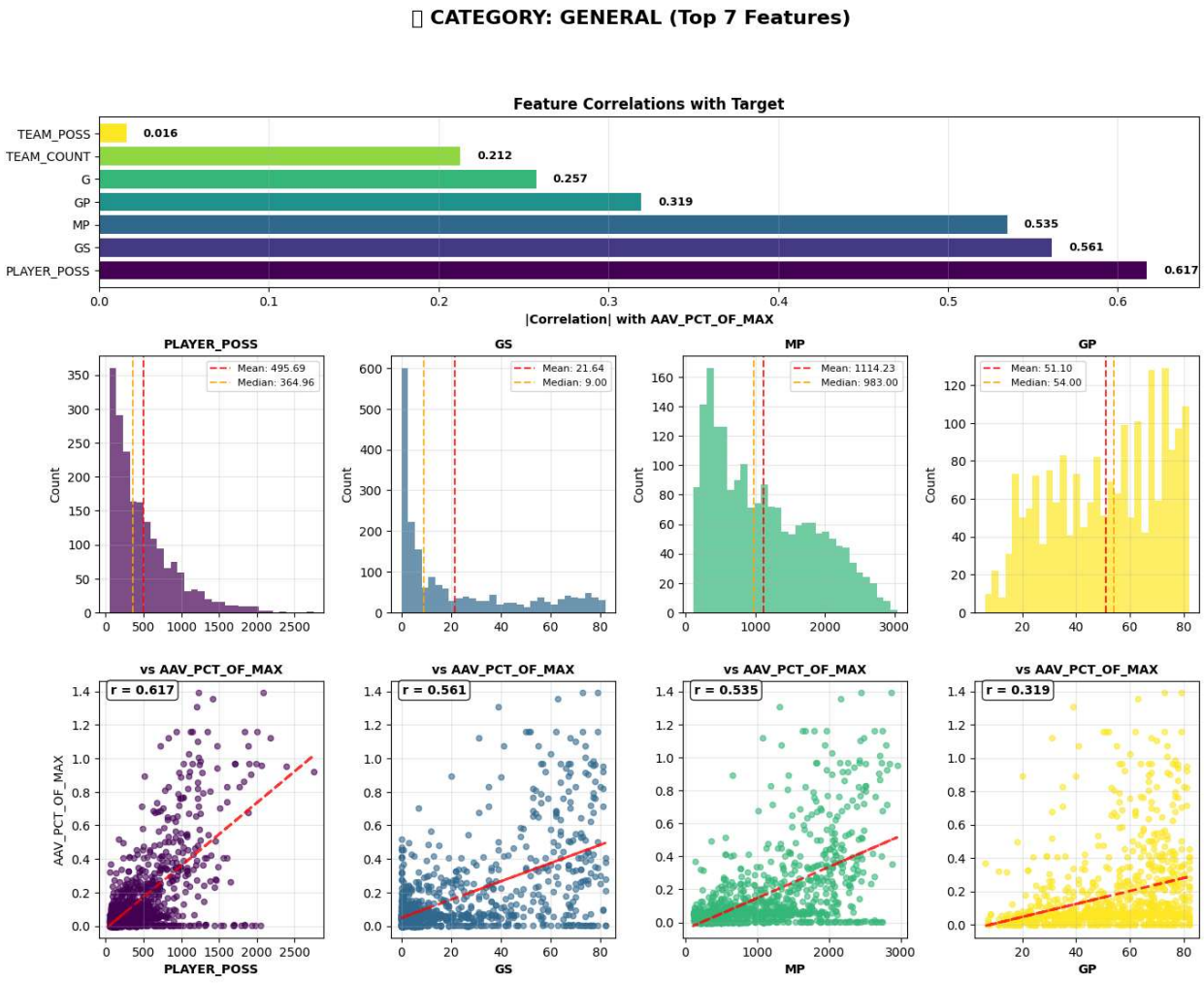
□ IMPORTANCE STATISTICS

Total Features Analyzed: 50
Features with Combined Score > 0.1: 50

□ METHOD AVERAGES:
Correlation: 0.491
Mutual Information: 0.515
Random Forest: 0.020

□ TOP FEATURE:
MIN_RANK
Combined Score: 0.973

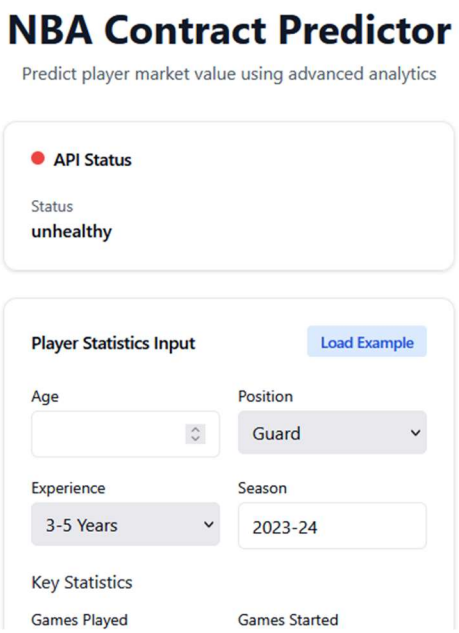
□ AGREEMENT BETWEEN METHODS:
Corr vs MI: 0.625
Corr vs RF: 0.146
MI vs RF: 0.384



Key findings summary:

The target variable AAV_PCT_OF_MAX is highly right-skewed (skewness 2.34, kurtosis 5.45) with a median of 0.067 and IQR [0.033–0.203], and about 10% of values flagged as potential outliers. Data quality is excellent—97.4% complete, 100% unique rows, and only 44 columns with >10% outliers. Feature importance analysis highlights contract-usage metrics (e.g. MIN_RANK, PTS_RANK, USG%) as strongest predictors. Among feature categories, General (e.g. PLAYER_POSS r = 0.617), Scoring (PTS r = 0.618), and Usage (USG%, TRUE_USAGE% r = 0.617) show the highest correlations with our target.

React Site Started:


NBA Contract Predictor
Predict player market value using advanced analytics
API Status
Status
unhealthy
Player Statistics Input
Load Example
Age
Position
Guard
Experience
3-5 Years
Season
2023-24
Key Statistics
Games Played
Games Started

---------- PROJECT 2----------