# Bike Sharing analysis Report

**Target Variable:** Target variable ‘cnt’ shows a strong right-skew with some high demand outliers, which are real demand spikes. Poisson GLM is a good fit because the target variable – count of bike rentals - is a non-negative count outcome. The Poisson GLM naturally models mean-variance relationship for count data, assumes the variance equals the mean, and uses a log link to ensure non-negative predictions. It provides interpretable coefficients for operational decision-making and serves as a standard baseline model to test for overdispersion. Other considerations like XGBoost or negative binomial later?

**Why normalize predictors?**

* Continuous predictors were normalized using max-min scaling to improve the numerical stability of the Poisson GLM fitting process, ensure faster convergence, and make coefficients more interpretable on a consistent scale.
* **Faster convergence**: Normalization improves convergence of iterative algorithms like IRLS in GLMs by ensuring predictors contribute proportionately to gradient and Hessian calculations. This prevents predictors with larger scales from dominating the step size, leading to faster, more stable convergence.
* Tree-based models like XGBoost do not require normalization, but consistent preprocessing was applied for comparability.

**Why remove certain variables?**

* Just an identifier – no predictive power.
* Target leakage results in artificially good results that won’t generalize.

**Univariate analysis:**

Various predictors are converted to factors for better analysis to ensure that the GLM treated them as categorical predictors. This allows the model to estimate separate effects for each category rather than assuming a numeric relationship.

There are no missing values.

All categorical variables were checked for sufficient representations:

* season, yr, mnth, hr, weekday show good balance across levels.
* Holiday occurs infrequently (~ 3% of observations) but is retained due to its operational importance.
* Weathersits shows category 4 (severe weather) with only 3 cases; this may not support stable coefficient estimates and could be combined with category 3 or dropped from the model. In this case we combined with cat 3.

All numeric predictors atemp, temp, humidity, windspeed were examined.

* All variables show reasonable spread within expected normalized ranges.
* Small humidity outliers near zero could be real (e.g. very dry days)
* Some high-end outliers in windspeed were identified; they will be retailed for now as they may represent real conditions but will be monitored for undue influence during model diagnostics like Cook’s distance, influential points.

**Relevel categorical variables:**

The baseline level of each categorical variable was set to the level with the largest number of observations. This ensures that the mean of the reference level is estimated with maximum precision, resulting in smaller standard errors for all level contrasts and more stable coefficient estimates.

**Multivariate analysis:**

Consider:

1. Which predictors have meaningful relationship with target
2. Detect possible nonlinearities
3. Detect possible interactions
4. Spot multicolieanrity

Numeric predictors vs numeric target:

Scatterplots with LOESS smoothing shows a clear positive trend between temperature and bike rentals, with a mild flattening at higher temperature. Humidity and windspeed both show mild negative relationship with rentals, suggesting less favourable conditions reduce demand slightly. These findings support the inclusion of temperature, humidity and windspeed as predictors in GLM, with temperature likely being the strongest driver.

Categorical predictors vs numeric target:

Bivariate analysis of hourly rentals by working day status shows distinct usage patterns:

Working day have sharp peaks during commute hours, while non-working days show a broader afternoon peak. This supports including both hour and working day as predictors and suggest an interaction between the two may improve the model.

Boxplot of bike rentals by season indicates significant seasonal variation. Fall, Summer and Winter show higher median and more variable rentals compared to Spring, which has the lowest median. This shows the importance of including seasonality in the model to account for predictive fluctuations in demand.

Multicollinearity:

Pairwise correlation analysis confirmed that temperature and apparent temperature are highly correlated (r = 0.988) indicating redundancy and VIF (Variance Inflation Factor) analysis identified strong multicollinearity between temp and atemp (VIF > 40. To avoid multicollinearity, only temp is retained for modeling. Other predictors show expected relationship with rentals, and no additional multicollinearity concerns were identified.