

# Capital Bikeshare Analytics Report

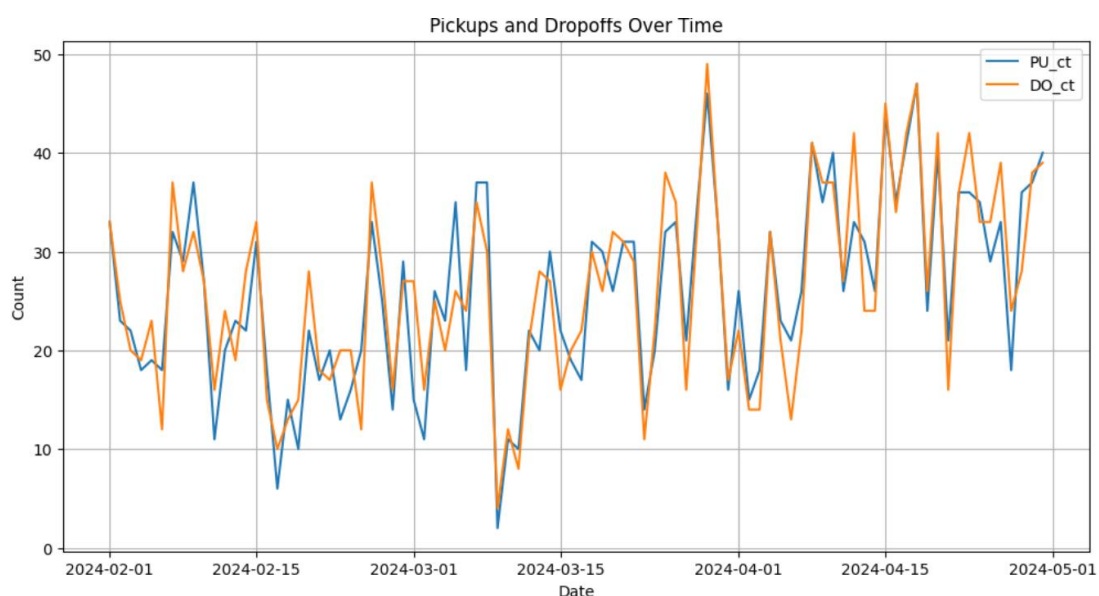
## 1. Business Understanding

The George Washington School of Business (GWSB) station at “**22<sup>nd</sup> & H St NW**” is frequently capacity-constrained. Our goal is to develop data-driven tools that helps in predicting *next-day* pickup (PU) and drop-off (DO) counts, translate those predictions into an *optimal lay-out* of bikes vs docks under different capacity scenarios, and design crew-routing clusters for the wider Capital Bikeshare network.

Success is judged along two axes. First **Prediction accuracy** – checking Mean-Squared-Error (MSE) on an out-of-sample test set and **Decision cost** – the penalty  $\alpha \cdot \text{missPU} + \beta \cdot \text{missDO}$  incurred by the station manager after allocating bikes/docks using the model’s forecast ( $\alpha = 2, \beta = 3$ ).

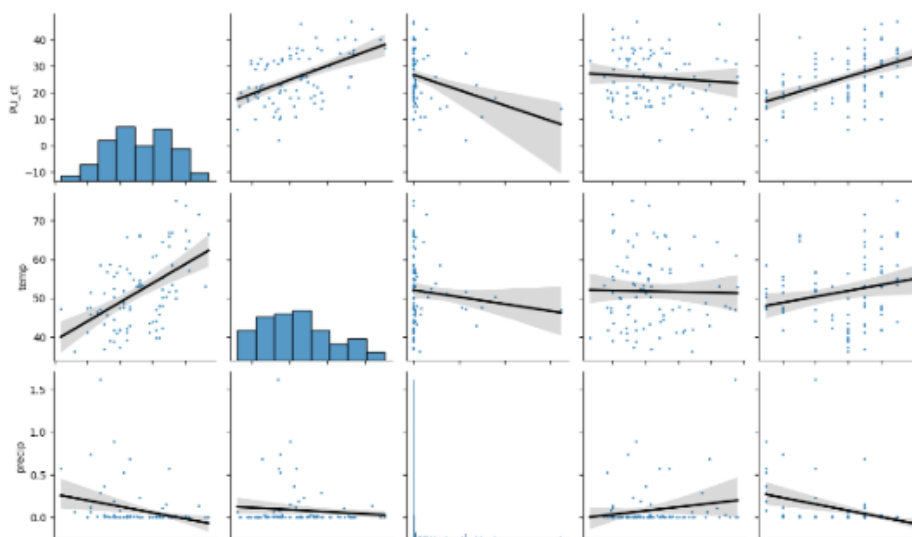
## 2. Exploratory Analysis

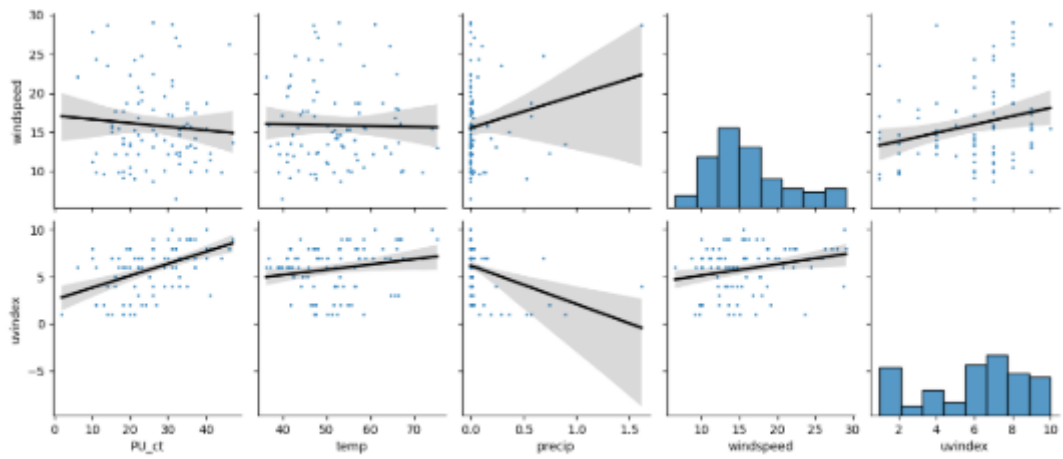
*Daily demand.* Between 1 Feb – 30 Apr the station handled **33 ± 8 pickups and 33 ± 7 drop-offs per day**. A clear weekend trough and weekday peak is visible.



**Figure 1: “Pick-ups and Drop-offs over Time” line chart**

*Weather joins.* We merged Visual Crossing’s 33-column DC weather feed with PU/DO counts by date. Pair-plots show a mild positive slope with temperature and a weak negative slope with precipitation & wind speed. UV-index correlates most strongly with both targets.





**Figure 2: 5 × 5 seaborn pair-plot**

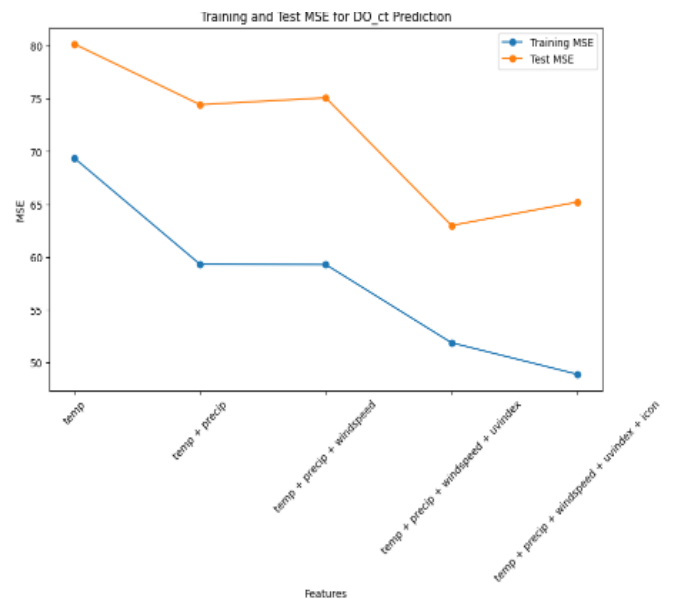
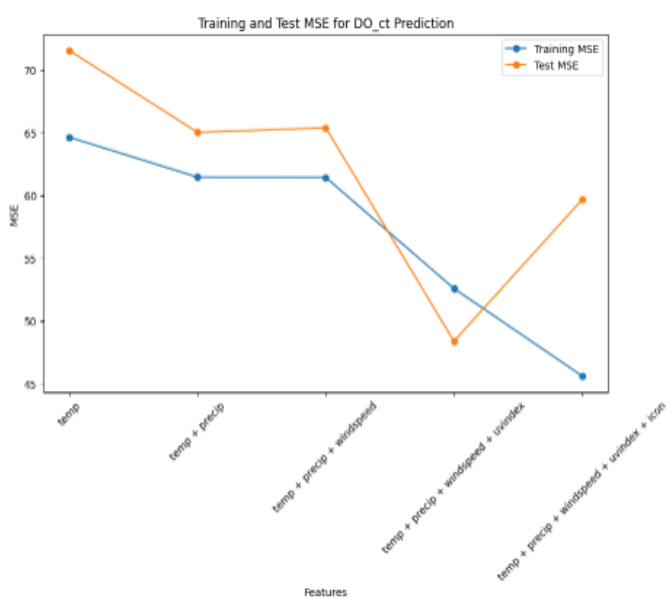
### 3. Predictive Modelling

#### 3.1 Linear-baseline ladder

Feature set	Train MSE	Test MSE
PU_ct – temp	64.63	71.54
temp + precip	61.45	65.02
temp + precip + windspeed	61.43	65.38
temp + precip + windspeed + uvindex	52.58	48.39
temp + precip + windspeed + uvindex + icon	45.62	59.66

**Table 1: Model performance with different features**

Identical patterns hold for DO\_ct (best test MSE = 62.99 at the 4-feature model).



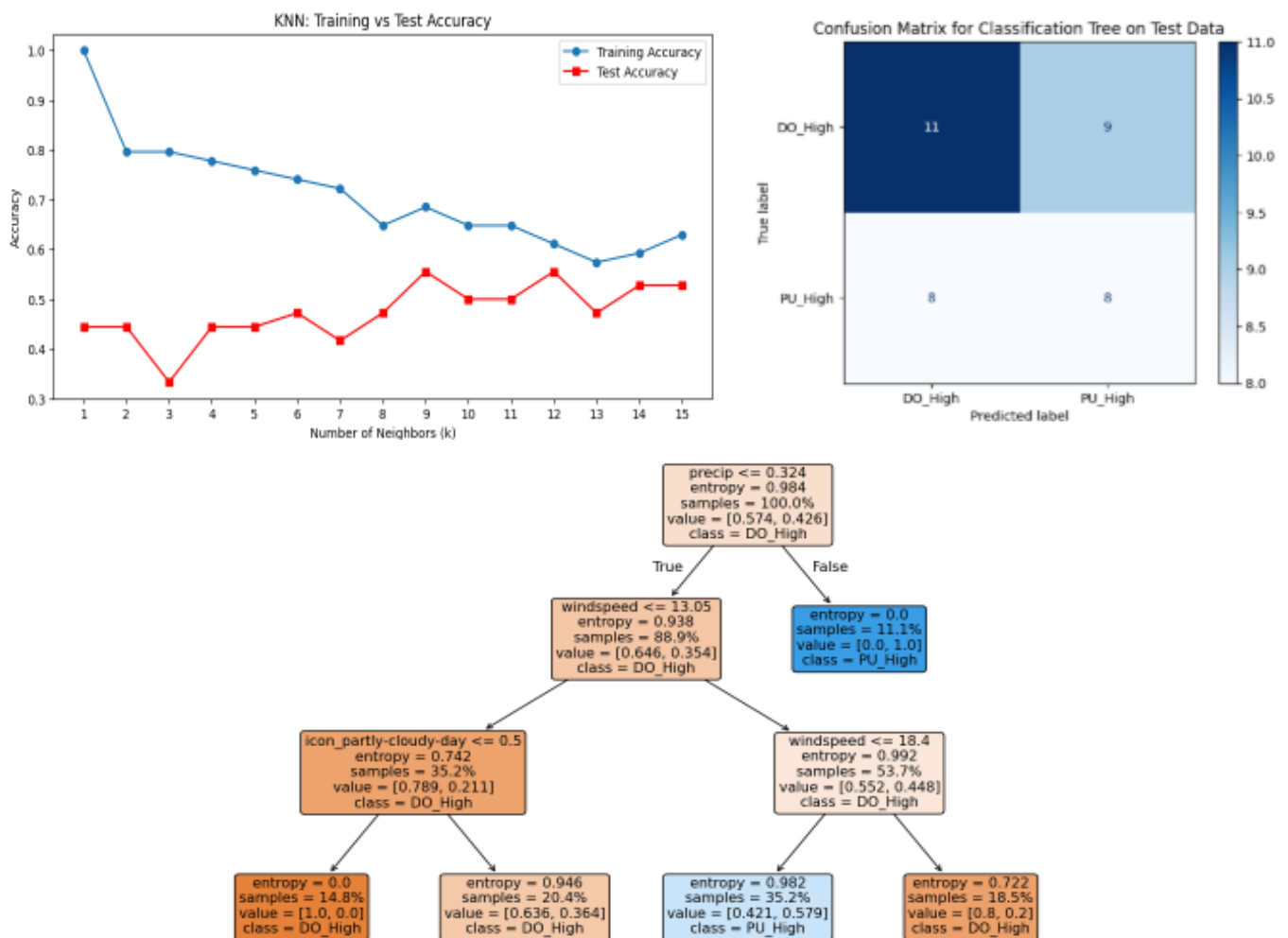
**Figure 3 & 4 – MSE-vs-features plots for PU\_ct & DO\_ct**

#### 3.2 Classification (PU > DO)

Five models classify whether pickups exceed drop-offs.

Model	Test accuracy
KNN (k = 9)	0.556
Logistic Reg	0.528
Linear SVC	0.500
RBF SVC	0.500
Tree (depth = 3)	0.528

**Table 2: Classification model and test accuracy**



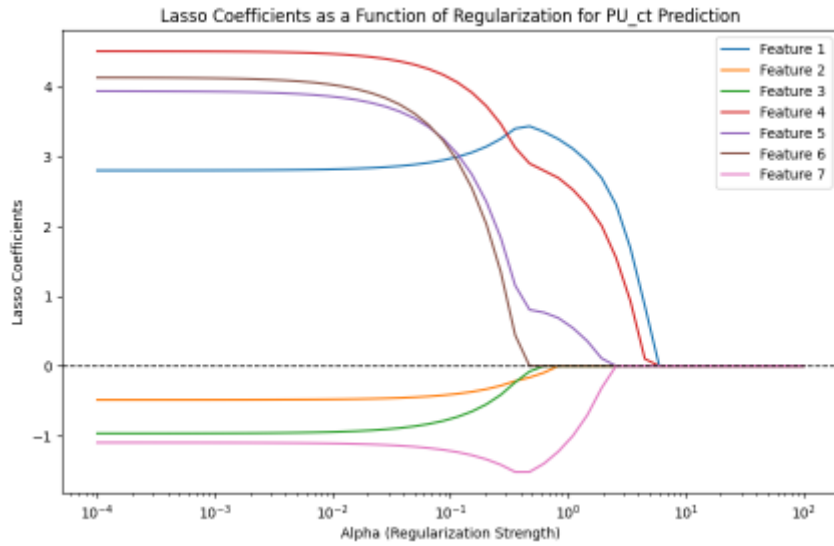
**Figure 5 – KNN accuracy-vs-k; Fig 6 – Confusion matrix; Fig 7 – Classification tree**

**Note:** TPR/FPR & ROC analyses treat **DO\_High** as the positive class throughout.

### 3.3 Regularisation (Lasso)

- PU\_ct: best  $\alpha = 0.83 \rightarrow$  CV MSE  $\approx 66.5$ , test MSE = 55.4
- DO\_ct: best  $\alpha = 0.63 \rightarrow$  test MSE = 68.6

Coefficient path (Fig 8) reveals that precip, windspeed, and the “rain” icon shrink to 0, leaving **temp, uvindex, sky-clear indicator & snow icon** as drivers.



**Figure 8 – Lasso coefficient-path**

### 3.4 Dimension-reduced, multi-output models (PCA → 4 PCs)

Nine algorithms were tuned via 100-draw RandomisedSearchCV (5-fold). Key results:

Model	Test MSE (PU)	Test MSE (DO)	Cost @K=20	Cost @K=30
Linear	<b>55.6</b>	<b>69.5</b>	70.14	50.14
Ridge	55.7	69.5	70.14	50.22
<b>GB</b>	75.7	97.0	<b>69.83</b>	<b>49.58</b>
Tree	81.5	98.8	69.92	50.75
RF	69.2	84.7	70.03	49.97
others	...	...	...	...

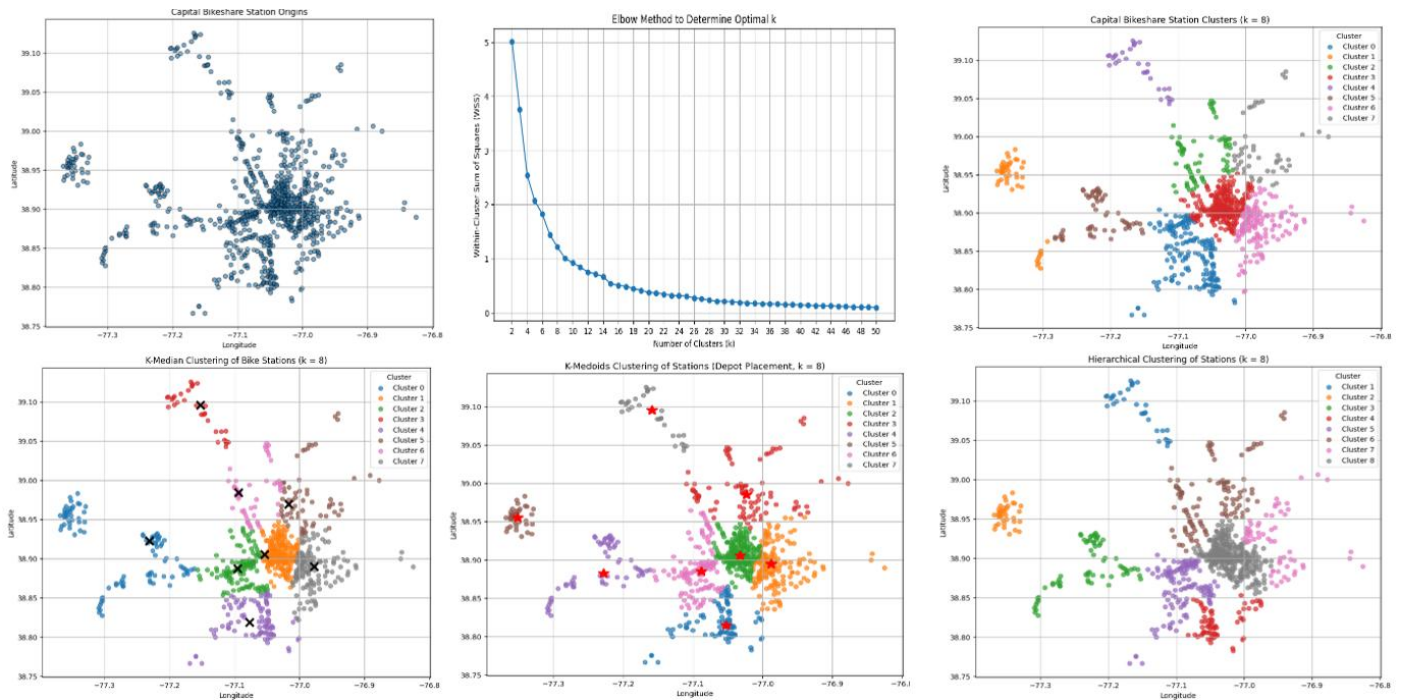
**Table 3: Classification model and test accuracy**

**Insight.** Gradient Boosting pays the *lowest operational cost* despite its mediocre MSE. Conversely, Linear/Ridge minimise prediction error but cost  $\approx 0.3$  units more per day. Choice should hinge on sponsor's tolerance for cost vs forecasting precision.

## 4. Clustering for Crew Deployment

Using all 880 unique stations (Fig 9):

1. **Elbow** suggests  $k \approx 8$  (Fig 10).
2. **k-means (k = 8)** – compact Euclidean spheres (Fig 11).
3. **k-median** (Manhattan) – grid-aligned blocks, robust to outliers (Fig 12).
4. **k-medoids** – chooses real stations as depots (Fig 13). Recommended depots listed below.
5. **Hierarchical (Ward, k = 8)** – irregular shapes, but reveals nested structure (Fig 14).



## 5. Performance Evaluation

### 5.1 Prediction metrics

- Lowest PU/DO MSE (simple features) → Linear/Ridge after PCA & scaling.
- Regularised Lasso discards noisy variables without hurting error.
- Classification accuracy peaks at ~56 % (still modest: data are noisy & balanced).

### 5.2 Decision metrics

- Under realistic capacity  $K = 20$ , Gradient Boosting lowers expected penalty by  $\approx 0.3$  units per day vs other models.
- At  $K = 30$  that lead widens (**49.6 vs 50.1**).

**Trade-off:** Because  $\alpha$ ,  $\beta$ , and  $K$  encode *business priorities*, model selection should be scenario-specific. If cost sensitivity trumps interpretability → deploy GB. If explainability & accuracy matter more → stick with Linear/Ridge and accept a tiny cost premium.

## 6. Conclusions & Recommendations

1. **Forecasting:** Use the 4-PC Linear model as the *default* forecaster; switch to GB during high-demand events when cost of shortages escalates.
2. **Real-time allocation:** Embed the decision-cost routine to suggest optimal bike/dock split each morning.
3. **Depot placement:** Adopt the 8 k-medoid depots for crew logistics; they minimise travel under Manhattan distances and are actual stations.
4. **Limitations:** Only 90 days of data – seasonality and special events not captured. Weather vs demand assumed stationary; sudden policy or pricing changes are ignored. Classification models under-perform; richer features (events, holidays) could help.
5. **Next steps.** Incorporate demand uncertainty (predictive distributions) and run Monte-Carlo optimisation; extend forecasting horizon to intra-day 15-minute intervals.