**Logistic Regression and PCA–K-means Analysis of Melbourne Housing Data**

## 1. Introduction

This report analyzes the Melbourne Housing dataset, which contains property-level information such as prices, structural characteristics, and location attributes.
The goal is twofold:

- Modeling whether a property sells at auction using a logistic regression model.
- Using PCA followed by K-means clustering to explore whether properties can be grouped based on key numerical characteristics.

These analyses provide complementary perspectives: the logistic regression focuses on prediction, while PCA + clustering focuses on unsupervised structure in the data.

## 2. Data Description

The subset of variables used in these analyses includes:

**For logistic regression:**

- **price**
- **rooms** – number of rooms
- **bathroom** – number of bathrooms
- **car** – number of parking spaces
- **distance** – distance from the Central Business District (CBD)
- **type** – h (house), t (townhouse), u (unit)
- **method** – we convert "S" to 1 (sold at auction day), all else to 0

Derived variables:

- **log_price** = log(price)
- **sold_auction** = 1 if method == "S", otherwise 0
- **pred_prob** = predicted probability from the logistic model

**For PCA and clustering:**

Numerical features:

- **rooms, bathroom, car, distance, landsize, building_area**, and others depending on availability
  All numeric variables were standardized before PCA.

## 3. Logistic Regression Model

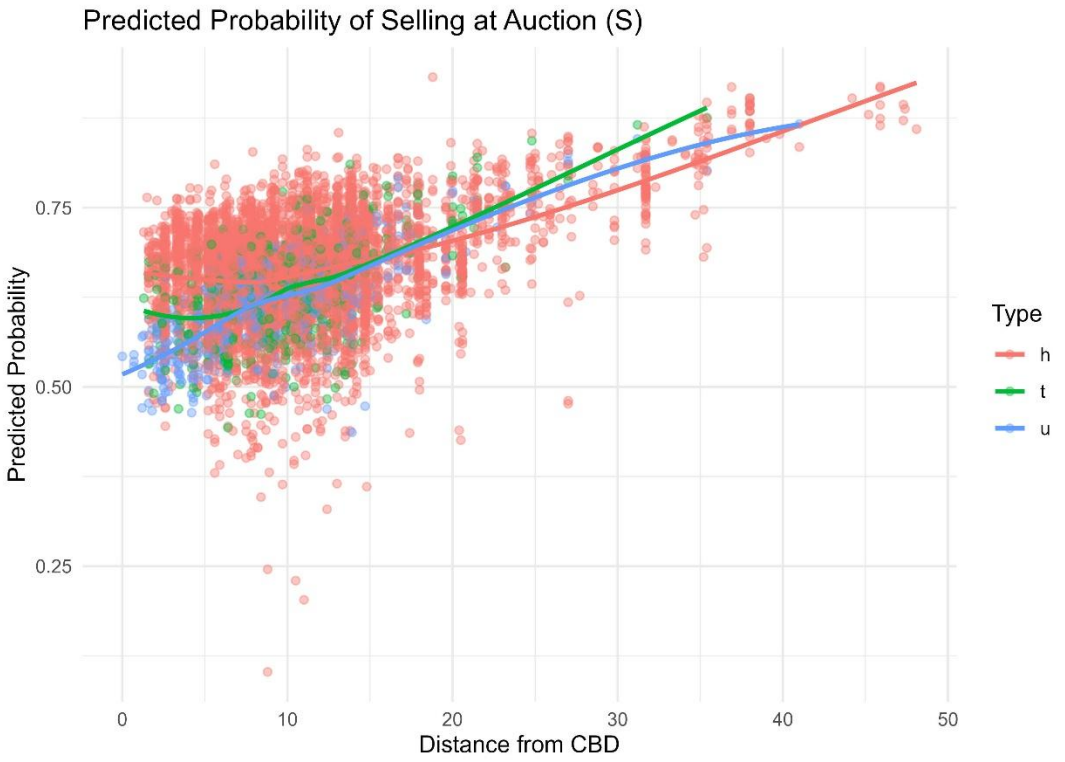
**The model predicts whether a property sells on the auction day:**

sold_auction ~ log_price + rooms + bathroom + car + distance + type

After fitting the model, predicted probabilities were calculated and visualized.

## 4. Logistic Regression Results

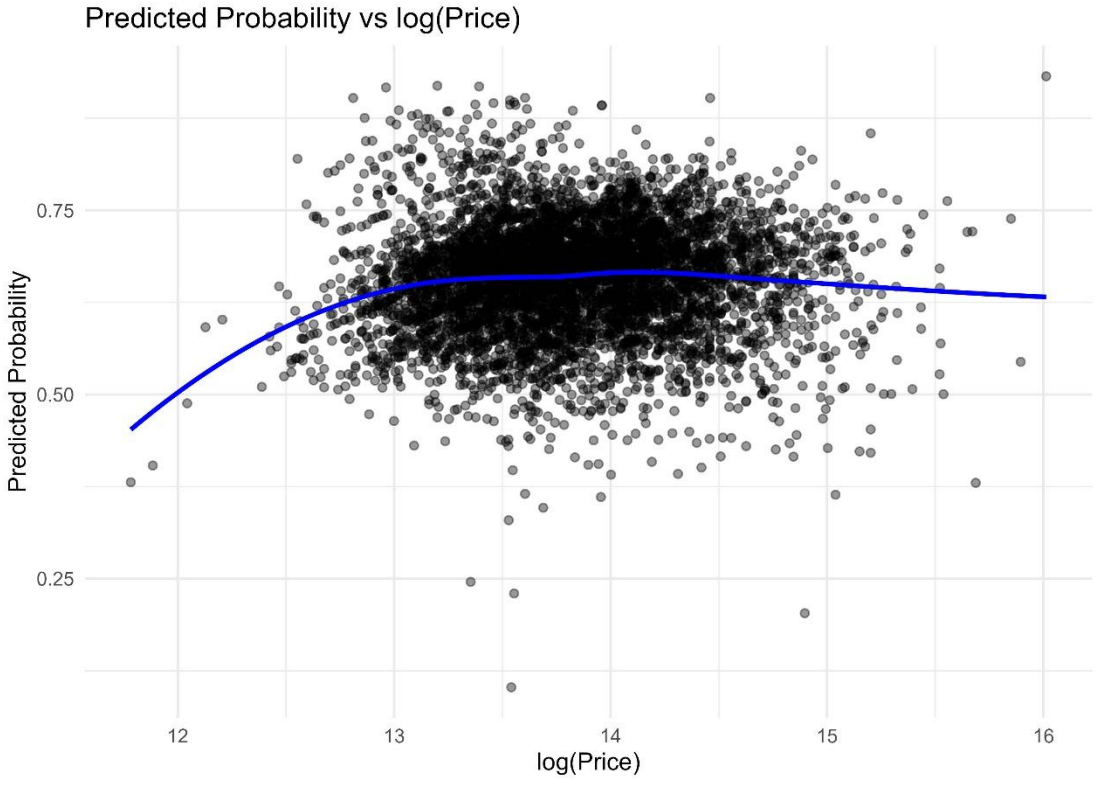
### 4.1 Predicted Probability vs Distance from CBD




**Interpretation:**

- Predicted probability of selling at auction increases with distance from the CBD.
- The trend holds across property types (h, t, u).
- This may reflect market differences between inner-city vs outer suburban areas.

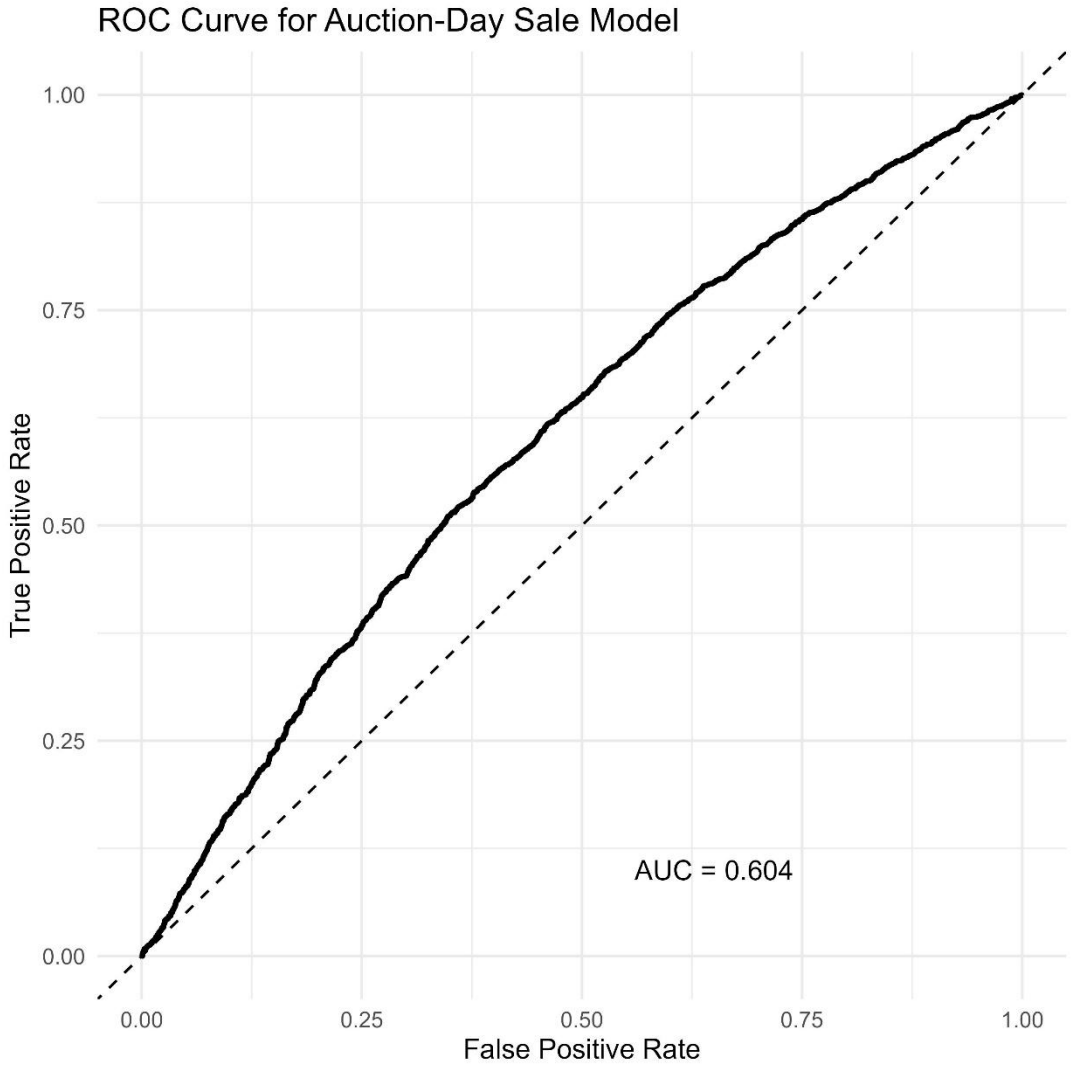
### 4.2 Predicted Probability vs Log(Price)



**Interpretation:**

- The relationship between price and auction success rate is non-linear.
- Probability increases at lower price ranges but levels off or slightly decreases at very high prices.
- This suggests that extremely expensive properties may rely less on auction-day purchases.

### 4.3 ROC Curve of the Logistic Model



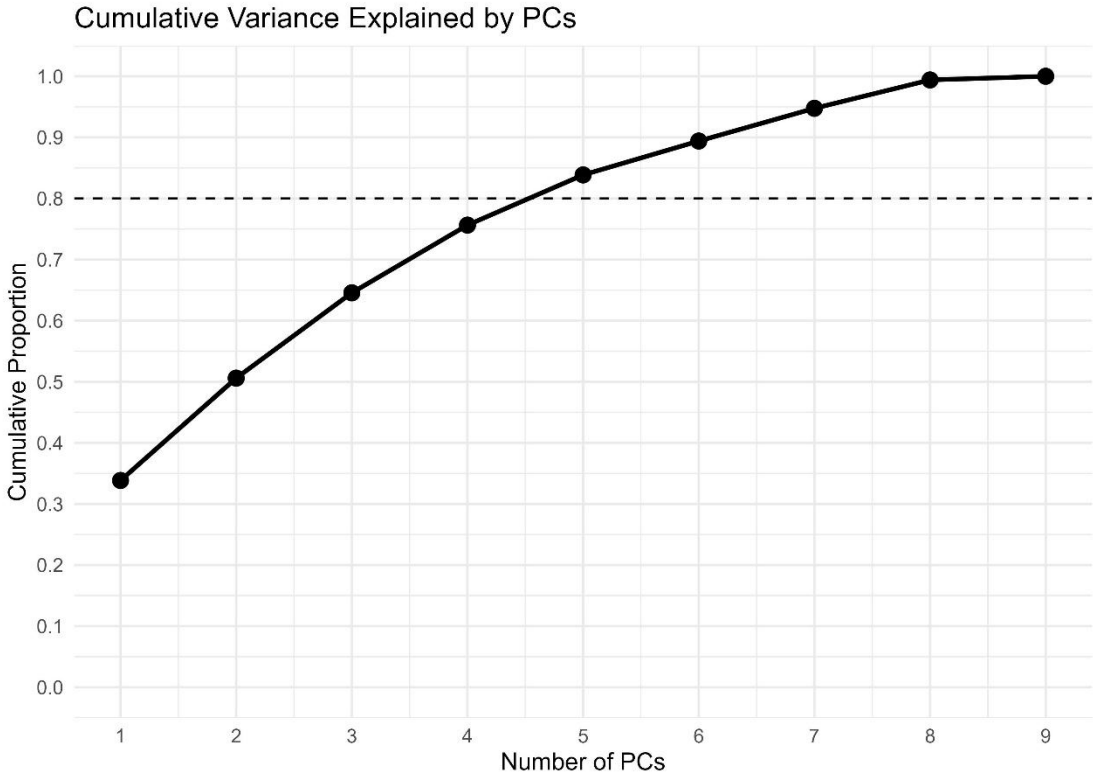

**AUC = 0.604**

**Interpretation:**

- An AUC of 0.604 indicates modest predictive performance.
- The model performs better than random guessing but leaves room for additional features or nonlinear modeling.

## 5. PCA and Clustering Analysis

### 5.1 Cumulative Variance Explained by PCs

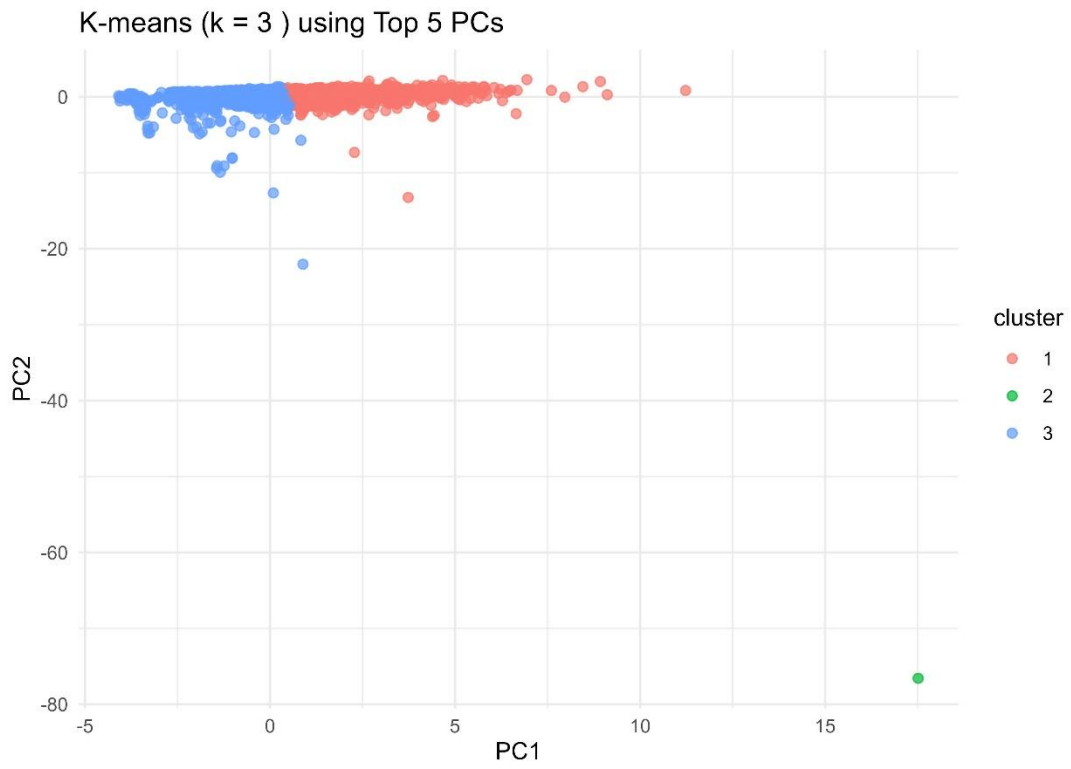



**Interpretation:**

- The first 5 principal components explain about 84% of total variance.
- Using the 5-PC representation is reasonable for clustering while maintaining most information.

### 5.2 K-means Clustering Using First 5 PCs

We apply K-means with k = 3, a common starting point.




**Interpretation:**

- With **k = 3 clusters**, the PCA space reveals **two large clusters** and **one very small cluster**.
- The extremely isolated points suggest:
  - Potential outliers (e.g., very large land size or building area)
  - Unique high-end or low-end properties not similar to the majority

**Implications:**

- Clustering does **not map cleanly** onto property types (h, t, u).
- PCA + K-means captures **structural differences**, but the dataset may require:
  - More features (e.g., location effects)
  - Pre-removal of extreme outliers
  - Alternative clustering methods (e.g., DBSCAN)

**6. Conclusion**

This report applied both supervised and unsupervised methods to the Melbourne Housing dataset.

**Key findings:**

- Logistic regression shows that distance from CBD and log(price) influence the probability of selling at auction.
- Model performance is moderate (AUC = 0.604), suggesting nonlinearity or missing predictors.
- PCA shows that 5 components explain ~84% of the variance.
- K-means clustering in the PCA space reveals clusters influenced by extreme or unusual properties rather than standard property types.

**Future improvements:**

- Incorporate spatial features (latitude, longitude)
- Consider non-linear models (Random Forest, GAM, Gradient Boosting)
- Investigate outliers before clustering
- Use domain-specific grouping methods