# Advanced data visualization

## Experiment-5

| Name | Hardik Garg |
|---|---|
| UID | 2021300036 |
| Batch | Batch H |
| Department | COMPS A |

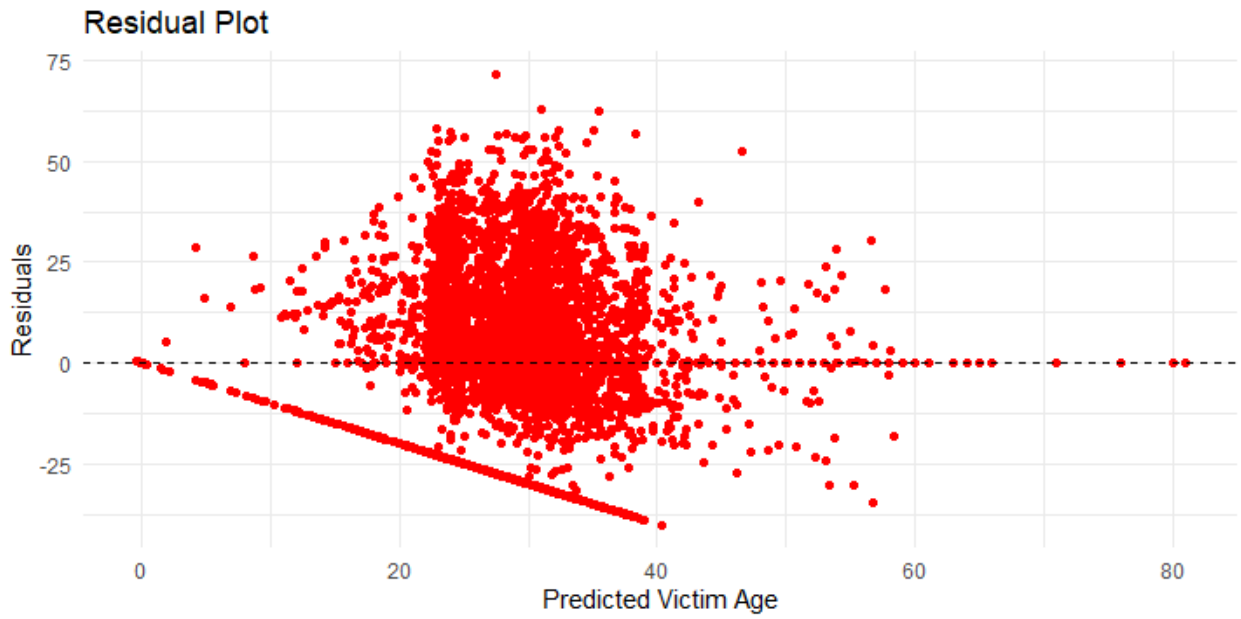
**Aim:** Create advance chart charts using PowerBi/Tableau/R/Python/D3.js on dataset Housing data set.

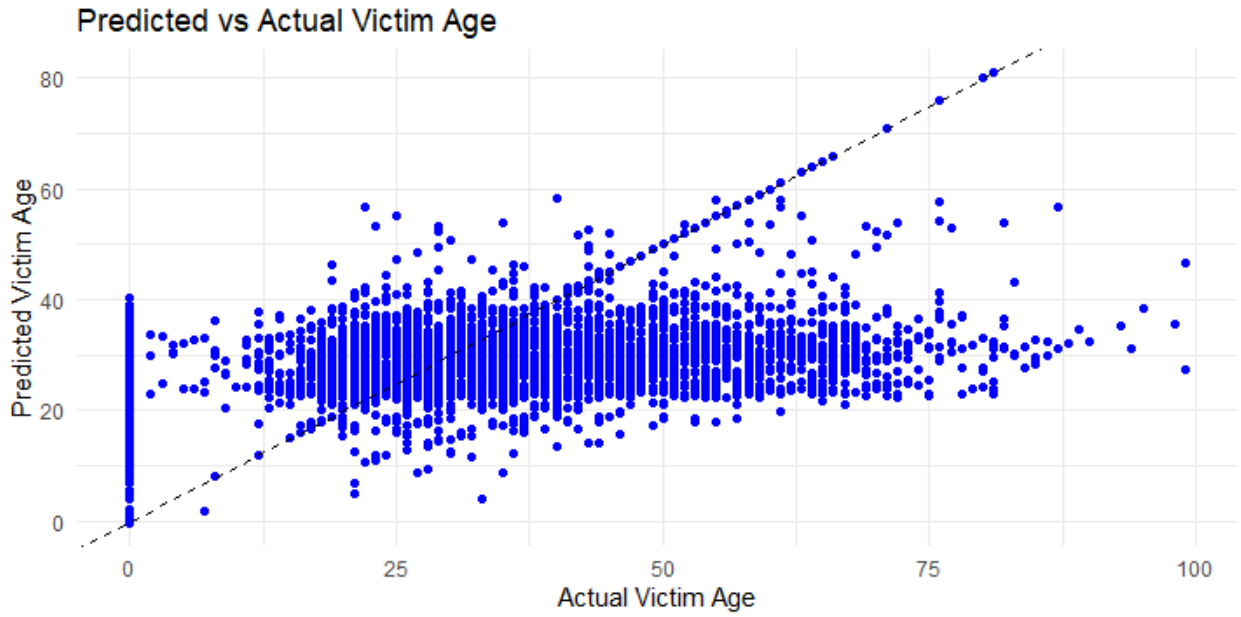
## Linear Regression

**Key Columns in the Dataset:**

**DR_NO**: *Integer* — Division of Records Number (unique identifier for each report).
**DATE.OCC**: *Character* — Date of occurrence (formatted as MM/DD/YYYY).
**TIME.OCC**: *Character* — Time of occurrence (formatted as HH
in 24-hour military time).
**AREA**: *Integer* — Geographic area number where the crime occurred (1-21).
**AREA.NAME**: *Character* — Name designation of the geographic area or patrol division.
**Rpt.Dist.No**: *Integer* — Reporting District Number (sub-area within the geographic area).
**Crm.Cd**: *Integer* — Crime Code (indicating the type of crime).
**Crm.Cd.Desc**: *Character* — Description of the Crime Code.
**Mocodes**: *Character* — Modus Operandi codes (methods or patterns associated with the crime).
**Vict.Age**: *Integer* — Age of the victim.
**Vict.Sex**: *Character* — Sex of the victim (F - Female, M - Male, X - Unknown).
**Vict.Descent**: *Character* — Descent of the victim (e.g., Hispanic, White, Black, etc.).
**Premis.Desc**: *Character* — Description of the premises where the crime occurred.
**Weapon.Used.Cd**: *Numeric* — Code representing the type of weapon used in the crime.
**Weapon.Desc**: *Character* — Description of the weapon used.
**Status**: *Character* — Status of the case (e.g., IC - Incomplete).
**Status.Desc**: *Character* — Description of the status.
**LOCATION**: *Character* — Street address of the crime incident, rounded to the nearest hundred block.
**LAT**: *Numeric* — Latitude coordinate of the crime location.
**LON**: *Numeric* — Longitude coordinate of the crime location.

```
# Set a seed for reproducibility (optional)
set.seed(123)

# Randomly sample 5000 rows from the dataset
sampled_data <- crime_data[sample(nrow(crime_data), 5000), ]

# Fit a linear regression model to predict Victim Age based on TIME.OCC and AREA
model <- lm(Vict.Age ~ TIME.OCC + AREA, data = sampled_data)

# Print the summary of the model to evaluate its performance
summary(model)

# Plot actual vs predicted Victim Age
library(ggplot2)

ggplot(sampled_data, aes(x = Vict.Age, y = predict(model))) +
  geom_point(color = 'blue') +
  geom_abline(slope = 1, intercept = 0, linetype = "dashed") +
  labs(x = "Actual Victim Age", y = "Predicted Victim Age", title = "Predicted vs Actual Victim
Age") +
  theme_minimal()

# Plot residuals to check model fit
ggplot(sampled_data, aes(x = predict(model), y = resid(model))) +
  geom_point(color = 'red') +
  geom_hline(yintercept = 0, linetype = "dashed") +
  labs(x = "Predicted Victim Age", y = "Residuals", title = "Residual Plot") +
  theme_minimal()
```




## Logistic Regression

**`price`**: Integer. The price of the property.

**`area`**: Integer. The size of the property in square units.

**`bedrooms`**: Integer. The number of bedrooms in the property.

**`bathrooms`**: Integer. The number of bathrooms in the property.

**stories**: Integer. The number of stories (floors) in the property.

**mainroad**: Character. Indicates whether the property is located on a main road ("yes" or "no").

**guestroom**: Character. Indicates whether the property has a guestroom ("yes" or "no").

**basement**: Character. Indicates whether the property has a basement ("yes" or "no").

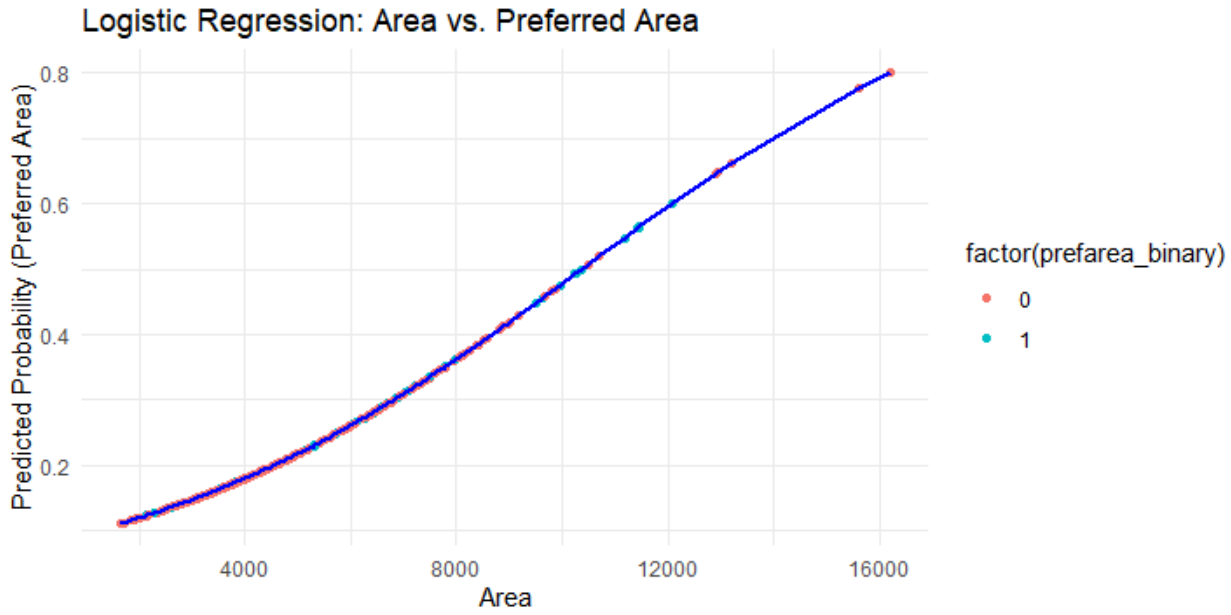
**hotwaterheating**: Character. Indicates whether the property has hot water heating ("yes" or "no").

**airconditioning**: Character. Indicates whether the property has air conditioning ("yes" or "no").

**parking**: Integer. The number of parking spaces available.

**prefarea**: Character. Indicates whether the property is in a preferred area ("yes" or "no").

**furnishingstatus**: Character. Indicates the furnishing status of the property (e.g., "furnished", "semi-furnished", "unfurnished").



**Conclusion-** In the logistic regression analysis, the focus is on predicting whether a property is in a preferred area based on its size. The sigmoid curve illustrates how the probability of a property being in a preferred area changes with size, helping to understand the likelihood of preference based on area.

The dataset contains detailed records of criminal incidents reported by the LAPD, including crime types, victim demographics, and crime locations. Key columns include crime codes, victim age and sex, and geographical coordinates. The data types range from numeric and integer to character, reflecting both categorical and continuous information. Initial analyses can focus on summarizing the distribution of crime types, examining patterns in victim demographics, and exploring geographic trends. This structured dataset provides a solid foundation for further statistical analysis and modeling to uncover patterns and insights related to crime occurrence and its influencing factors.