UNIT–5 : Visualization (PYQ Solution)

PART–A Answers (≈70 Words Each)

Q1. What is a box plot? What does IQR mean? (Expected)

A box plot (box-and-whisker plot) displays the five-number summary of a numeric variable: minimum, first quartile Q1, median, third quartile Q3 and maximum. The box spans Q1 to Q3, a line marks the median, whiskers extend to the non-outlier extremes and dots show outliers.

IQR (interquartile range) $= Q3 - Q1$ is the spread of the middle 50 percent of the data. Values below $Q1 - 1.5\,IQR$ or above $Q3 + 1.5\,IQR$ are treated as outliers.

Q2. What is a heat map? (Expected)

A heat map is a grid-based chart in which the colour of each cell represents its value, so patterns are seen at a glance. Warm colours usually mean high values and cool colours low values.

It is widely used to show a correlation matrix, missing values, or the density of events. In R: heatmap(cor(mtcars)) or ggplot2 with geom_tile().

Q3. What is data balancing? Why is it needed? (Expected)

Data balancing means adjusting a dataset in which one class is much rarer than another (class imbalance), using oversampling, undersampling or SMOTE.

It is needed because a model trained on imbalanced data favours the majority class. For example, with 950 normal and 50 fraud records, always predicting "normal" gives 95 percent accuracy but detects no fraud.

PART–B Answers (≈250 Words Each)

Q1. Differentiate between histogram and scatter plot. Write R code for both. (Expected)

	Histogram	Scatter plot
Variables	One numeric variable	Two numeric variables
Shows	Distribution: shape, skewness, spread, modes	Relationship: trend, correlation, clusters, outliers
Axes	x = bins, y = frequency	x = one variable, y = another
Marks	Adjacent bars	Individual points

A **bar chart** is different from a histogram: bars represent categories, not continuous bins, and have gaps.

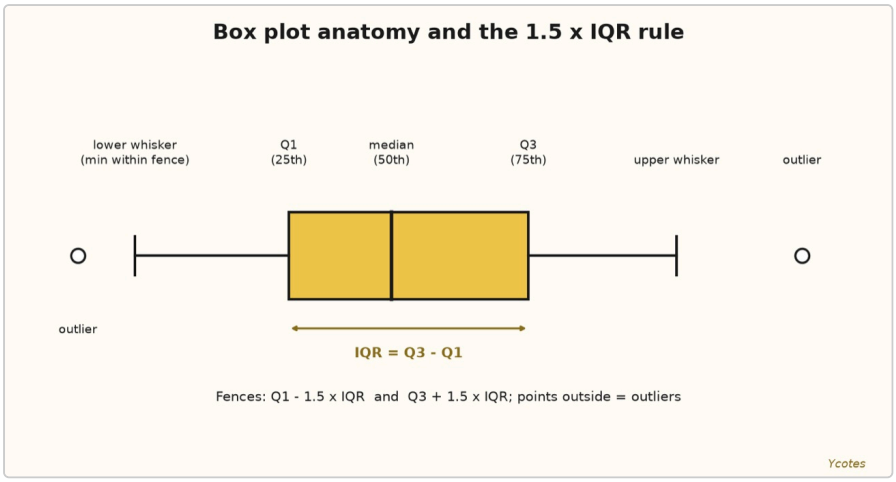
```
# Base R
hist(df[["marks"]], breaks = 10, col = "gold", main = "Marks", xlab = "Marks")
plot(df[["hours"]], df[["marks"]], pch = 19, xlab = "Hours", ylab = "Marks")

# ggplot2
library(ggplot2)
ggplot(df, aes(x = marks)) + geom_histogram(bins = 10, fill = "gold", colour = "black")
ggplot(df, aes(x = hours, y = marks)) + geom_point() + geom_smooth(method = "lm")
```

Reading them: a right-skewed histogram has a long tail to the right; an upward cloud in a scatter plot means positive correlation, which can be confirmed with cor(hours, marks).

Q2. Explain outlier detection using the IQR rule and the z-score with a

numerical. (Expected)



An **outlier** is a value that is very different from the rest of the data. It may be an error or a genuine rare event, and it can distort the mean, standard deviation and models.

Data: 12, 15, 14, 10, 18, 16, 13, 95, 11, 17 (n = 10). Sorted: 10, 11, 12, 13, 14, 15, 16, 17, 18, 95.

IQR rule:
$Q1 = 12.25$, $Q3 = 16.75$, $IQR = 16.75 - 12.25 = 4.5$
Lower fence $= 12.25 - 1.5(4.5) = 5.5$; upper fence $= 16.75 + 1.5(4.5) = 23.5$
95 lies above 23.5, so **95 is an outlier**; all other values are inside the fences.

Z-score rule: $z = (x - \mu)/\sigma$ with cut-off $|z| > 3$.
$\mu = 221/10 = 22.1$, $\sigma = 25.74$ (sample), so $z(95) = (95 - 22.1)/25.74 = 2.83$.
This is below 3, so the z-score rule **fails to flag** it. The outlier itself inflates μ and σ (masking). Without 95, $\mu = 14$ and $\sigma = 2.74$, and the z-score of 95 would be about 29.6.

Conclusion: the IQR rule is robust because quartiles are not affected by extreme values; use it for small or skewed data.

```
q <- quantile(x, c(0.25, 0.75)); i <- IQR(x)
x[x < q[1] - 1.5 * i | x > q[2] + 1.5 * i]      # 95
boxplot.stats(x)[["out"]]                        # 95
```

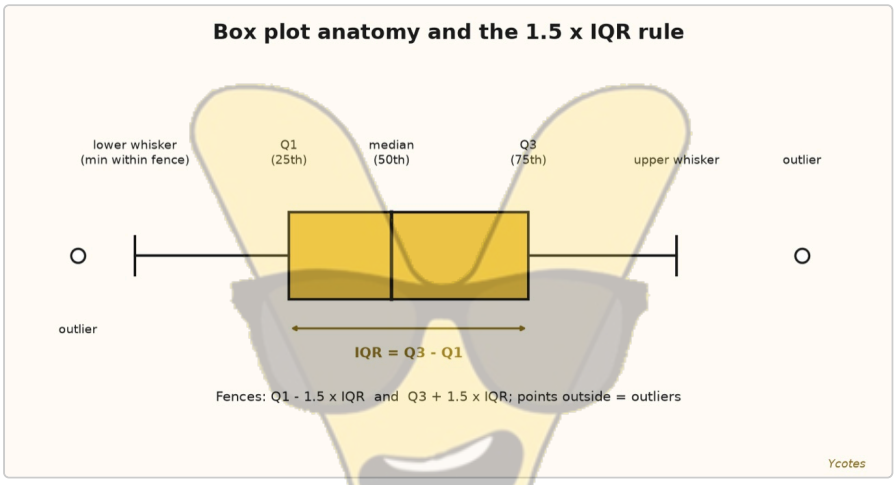
Treatment: correct if an error, remove, cap (winsorise) or transform (log).

PART–C Answers (≈500 Words Each)

Q1. Explain the visualization techniques box plot, histogram, scatter plot and heat map with R (ggplot2) code. (Expected)

Visualization turns data into pictures to explore patterns (Explore step of OSEMN) and to communicate results. **ggplot2** builds a plot from layers: data, aesthetics (aes), geometry (geom_), scales, facets and theme, joined with +.

1. Box plot: compares the distribution of a numeric variable across groups and shows outliers.




```
ggplot(mtcars, aes(x = factor(cyl), y = mpg)) + geom_boxplot(fill = "gold")
```

2. Histogram: distribution of one numeric variable.

```
ggplot(mtcars, aes(x = mpg)) + geom_histogram(bins = 10, fill = "gold", colour = "black")
```

3. Scatter plot: relationship between two numeric variables; colour or size can add a third.

```
ggplot(mtcars, aes(x = wt, y = mpg, colour = factor(cyl))) + geom_point(size = 3) +
geom_smooth(method = "lm", se = FALSE)
```

4. Heat map: colour-coded matrix, mostly for correlations.

```
cm <- round(cor(mtcars), 2)
heatmap(cm)                                    # base R
library(reshape2)
ggplot(melt(cm), aes(Var1, Var2, fill = value)) + geom_tile() + scale_fill_gradient2()
```

Summary:

Plot	Variables	Best for	geom
Box plot	numeric by category	spread, median, outliers	geom_boxplot
Histogram	1 numeric	distribution shape	geom_histogram
Scatter	2 numeric	correlation, clusters	geom_point
Heat map	matrix	correlation, density	geom_tile

Good practice: label axes, add a title, avoid clutter, use colour-blind-safe palettes and choose the chart by the question asked.

Q2. Explain data balancing techniques (oversampling, undersampling, SMOTE) and give an introduction to Tableau. (Expected)

Class imbalance occurs when one class has far fewer records, for example 950 non-fraud and 50 fraud (ratio 19:1). The model learns mostly the majority class, so accuracy looks high while recall for the minority class is poor.

Technique	Working	Drawback
Random undersampling	Remove majority rows (950 reduced to 50)	Throws away useful data
Random oversampling	Duplicate minority rows (50 raised to 950)	Overfitting to repeated points
SMOTE	Create new synthetic minority points on the line between a minority point and one of its k nearest minority neighbours	Can add noise near the boundary
Class weights	Give minority errors a higher cost in the loss	Needs support in the model

```
library(ROSE)
bal <- ovun.sample(Class ~ ., data = train, method = "both", N = 1000, seed = 1)
[["data"]]
table(bal[["Class"]])
```

Rules: balance only the **training** data, never the test data, and evaluate with precision, recall, F1 or AUC.

Tableau: a drag-and-drop business-intelligence tool for interactive visual analytics without coding.

1. **Connect** to data: Excel, CSV, SQL databases, cloud sources.
2. **Dimensions and measures:** dimensions are categorical fields (blue pills), measures are numeric fields (green pills).
3. **Worksheet:** drag fields to Rows and Columns; Tableau picks a chart (bar, line, map, scatter, heat map).
4. **Filters, sort, calculated fields, parameters** to refine the view.
5. **Dashboard:** combine sheets into an interactive dashboard and **publish** to Tableau Server, Cloud or Public.

Tableau is used for presenting results (iNterpret step), while R is used for modelling; the two can be connected through Rserve.

