UNIT–4 : BI – Data Analytics Process (PYQ Solution)

AKTU · Business Intelligence and Analytics (BCDS051) · Solutions by **Chirag Jain**, Ycotes

SECTION A Answers (≈70 Words Each)

Q1. Define evolution and deviation analysis. (Jan 2023)

Evolution analysis studies how data objects change over time, describing trends and regularities (for example, yearly growth of sales or stock price trends). Deviation analysis finds differences between actual and expected values (outliers, unusual changes), such as sales that fall far below the forecast, helping detect anomalies and fraud.

Q2. What are the 4 types of business analytics? (Jan 2023)

1. **Descriptive analytics: what happened (reports, dashboards).**
2. **Diagnostic analytics: why it happened (drill-down, root cause).**
3. **Predictive analytics: what is likely to happen (forecasting, machine learning).**
4. **Prescriptive analytics: what should be done (optimisation, recommendations).**

Q3. What are the types of variables in a dataset? (Jan 2023)

Variables are mainly of two kinds. Categorical (qualitative): nominal (no order, e.g. gender, colour) and ordinal (ordered, e.g. rating low/medium/high). Numerical (quantitative): discrete (counts, e.g. number of children) and continuous (measured, e.g. height, temperature), further interval and ratio scales. Variables may also be independent (predictor) or dependent (outcome).

SECTION B Answers (≈250 Words Each)

Q1. Explain Monte Carlo and regression analysis in detail. (Jan 2023)

Monte Carlo analysis (simulation) is a technique that uses **repeated random sampling** to model the probability of different outcomes in a process with uncertainty.

Steps:

1. Define the model and the output formula (e.g. profit = revenue − cost).
2. Assign **probability distributions** to uncertain inputs (demand, price, cost).
3. Generate random values for the inputs and compute the output.
4. Repeat thousands of times.
5. Analyse the distribution of results: mean, standard deviation, probability of loss, confidence intervals.

Uses: risk analysis, project cost and schedule estimation, financial forecasting, inventory planning. **Advantages:** handles uncertainty and many variables, gives a range, not a single value. **Limitation:** needs good distribution estimates and many runs.

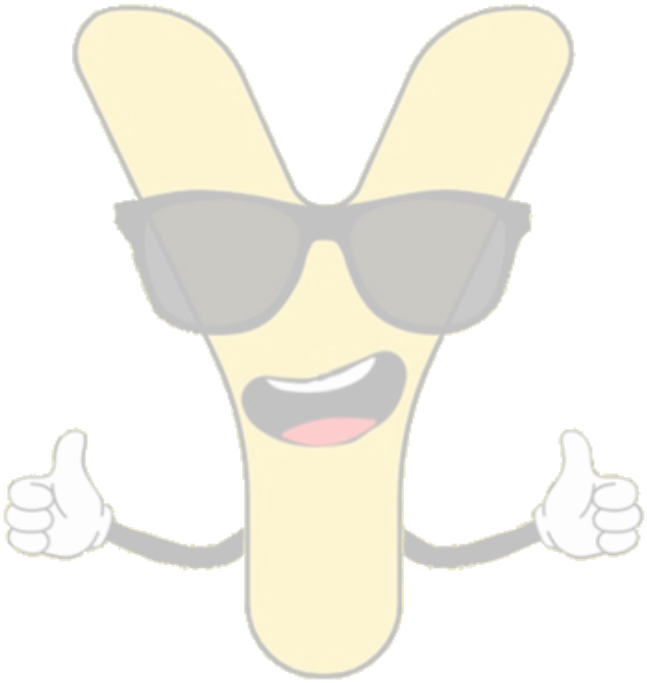
Regression analysis is a statistical method to find the **relationship between a dependent variable (Y) and one or more independent variables (X)** and to predict Y.

- **Simple linear:** $Y = a + bX + e$, where b is the slope and a is the intercept. Estimated by **least squares**, minimising $\Sigma(Y - \hat{Y})^2$.
- **Multiple linear:** $Y = a + b_1X_1 + b_2X_2 + \ldots + e$.
- Other forms: **logistic** (binary outcomes), polynomial, ridge/lasso.
- **Evaluation:** R^2 (fit), p-values, residual analysis, RMSE.

Example: predicting sales from advertising spend; slope shows extra sales per unit of spend.

Basis	Monte Carlo	Regression
Purpose	Simulate uncertainty and risk	Model relationships and predict
Method	Random sampling	Curve fitting
Output	Distribution of outcomes	Equation and predictions

Both are used together: regression builds the model, Monte Carlo tests it under uncertainty.



SECTION C Answers (≈500 Words Each)

Q1. How do you determine the accuracy of a model on the iris dataset? (Jan 2023)

The **Iris dataset** has 150 flowers, 4 features (sepal and petal length and width) and 3 classes (setosa, versicolor, virginica). Accuracy of a classifier is found as follows.

1. **Load and prepare data:** check for missing values, scale features if needed.
2. **Split** data into training and test sets (for example 80:20, stratified so classes stay balanced).
3. **Train** a model (k-NN, decision tree, logistic regression, SVM) on the training set.
4. **Predict** the classes of the test set.
5. **Compare** predictions with actual labels.

Accuracy = (correct predictions) / (total predictions). Example: 29 correct out of 30 test flowers = 96.7%.

Confusion matrix (3 × 3): rows are actual and columns predicted classes; the diagonal shows correct predictions. From it compute:

- **Precision** = TP / (TP + FP)
- **Recall** = TP / (TP + FN)
- **F1-score** = 2PR / (P + R)

Better validation: use **k-fold cross-validation** (for example 10-fold): train on k−1 parts, test on the remaining one, repeat k times and average the accuracy. This avoids dependence on one lucky split.

```
from sklearn.datasets import load_iris
from sklearn.model_selection import train_test_split
from sklearn.neighbors import KNeighborsClassifier
from sklearn.metrics import accuracy_score, confusion_matrix
X, y = load_iris(return_X_y=True)
Xtr, Xte, ytr, yte = train_test_split(X, y, test_size=0.2, stratify=y, random_state=1)
pred = KNeighborsClassifier(3).fit(Xtr, ytr).predict(Xte)
print(accuracy_score(yte, pred)); print(confusion_matrix(yte, pred))
```

Typical accuracy on iris is 95-98%. Setosa is perfectly separable; the confusion arises between versicolor and virginica.
