Final Exam Practice – COMP3602

Part 1: Introduction to Data Analysis & Visualization (L1)

1. What is the difference between data, information, and insight?
2. What are the main steps in a typical data analysis process?
3. List and explain the four types of data analysis.
4. Define data visualization and its importance.
5. Differentiate between descriptive and predictive analysis.
6. What is the role of data science in decision-making?
7. Mention three tools used in data analysis.
8. What is the difference between structured and unstructured data?
9. Give examples of binary, nominal, and ordinal data.
10. Explain the term “data storytelling.”
11. What are the benefits of effective data visualization?
12. Name three Python libraries used in data analysis or visualization.
13. What is the difference between univariate and multivariate data?
14. How does data quality affect analysis results?
15. What is the difference between static and dynamic data?

Part 2: Data Collection & Sampling (L2)

1. What is the importance of data collection in analysis?
2. Define population and sample in the context of statistics.
3. What makes a data sample representative?
4. List the types of sampling methods.
5. What is the difference between simple random and stratified sampling?

6. Describe cluster sampling with an example.
7. What is the purpose of cross-validation?
8. Define sampling bias and give an example.
9. What are common sources of sampling errors?
10. When is convenience sampling used and what are its risks?
11. What are the differences between primary and secondary data sources?
12. What kind of sample was used in the Shura Council election in Oman?
13. What does a sampling frame refer to?
14. Differentiate between quota and judgmental sampling.
15. What is a non-response error?

Part 3: Data Presentation (W3)

1. What are the main methods of data presentation?
2. What is a frequency distribution?
3. Differentiate between qualitative and quantitative frequency tables.
4. Define the following terms: class, class boundaries, and class mark.
5. How do you calculate class width?
6. What is the relative frequency and how is it computed?
7. What is the difference between a bar chart and a histogram?
8. When should a pie chart be used instead of a bar chart?
9. What is the purpose of an ogive curve?
10. What does a frequency polygon represent?
11. How are class intervals determined?
12. What is the difference between cumulative and non-cumulative frequency?
13. Which types of charts are best for categorical data?
14. What is a boxplot used for?

15. How can visualizations help detect outliers?

Part 4: Data Cleaning (W4)

1. What are common issues in raw data?
2. List four indicators of good data quality.
3. What are the main steps of data cleaning?
4. Provide two reasons for missing data.
5. List different methods for handling missing values.
6. Write Python code to replace missing values with the mean.
7. What is data noise and how can it be reduced?
8. What is the difference between binning and smoothing?
9. Explain how clustering helps in identifying outliers.
10. What are the consequences of having duplicate records in your dataset?
11. Give an example of an inconsistency that may occur when merging datasets.
12. What is a good strategy when dealing with over 50% missing values in a column?
13. What Python function is used to drop rows with null values?
14. How does regression help in smoothing noisy data?
15. Why is cleaning data critical before training a machine learning model?

Part 5: Data Exploration & Visualization (L5)

1. What is univariate analysis?
2. List three measures of central tendency.
3. How is median calculated for an even number of values?
4. What is the trimmed mean and why is it used?
5. Define standard deviation and variance.
6. What is the formula for Interquartile Range (IQR)?

7. What are the characteristics of a normal distribution?
8. How do you detect skewness in data using a boxplot?
9. Define multimodal distribution.
10. What is correlation and how is it different from covariance?
11. What does a Pearson correlation coefficient of 0.85 indicate?
12. What is a heatmap used for?
13. What is the purpose of a covariance matrix?
14. When should a frequency polygon be used instead of a histogram?
15. Explain the difference between lossy and lossless visualizations.

Part 6: Data Transformation & Reduction (L6)

1. What is the difference between Min-Max and Z-score normalization?
2. Apply Min-Max normalization to value 30 within the range [10, 50].
3. When is Decimal Scaling normalization appropriate? Give an example.
4. What is the goal of Discretization in data preprocessing?
5. Differentiate between Aggregation and Attribute Construction.
6. Explain the difference between Parametric and Non-parametric Reduction.
7. When and why do we use PCA in data preprocessing?
8. Why is Dimensionality Reduction important for machine learning?
9. What does Standardization produce in terms of data distribution?
10. Why is Min-Max normalization not ideal with outliers?
11. Differentiate between Normalization and Data Compression.
12. Provide a practical example of aggregation in real data.
13. Why is Attribute Construction crucial in predictive modeling?
14. Contrast Data Transformation with Data Reduction.
15. What type of reduction is applied when representing data using regression?

Part 7: Regression Analysis (L7)

1. What is the main goal of regression analysis?
2. Differentiate between simple and multiple linear regression.
3. Provide the general formula for simple linear regression.
4. State the formula to calculate the slope (B1) in regression.
5. What does a negative slope (B1 = -2) imply about the relationship?
6. Interpret $R^2 = 0.9$ in a regression model.
7. What does MSE represent and why is it used?
8. How does MSE differ from R^2?
9. Provide a real-world example where linear regression is applicable.
10. What is the purpose of sklearn's LinearRegression()?
11. Differentiate between the coefficient and intercept.
12. Can regression handle categorical data directly?
13. How can regression explain the relationship between price and quality?
14. Why should outliers be removed before regression?
15. Is linear regression suitable for non-linear relationships?
16. What is a limitation of using R^2 alone as a model metric?
17. What does an inverse relationship in regression indicate?
18. How do outliers affect the regression line?
19. How does logistic regression differ from linear regression?
20. Why include multiple independent variables in regression analysis?

Part 8: Time Series Analysis (L8)

1. How do time series data differ from regular tabular data?
2. List the three key characteristics of a stationary time series.
3. How can you detect if a time series is non-stationary?

4. Define trend and seasonality in a time series.
5. Provide an example of seasonality in real data.
6. Why do we need stationary series before applying ARMA?
7. Differentiate between AR and MA components in time series models.
8. What are the components of an ARIMA model?
9. Interpret the ARIMA(1,1,1) configuration.
10. Contrast Differencing and Detrending.
11. What is the purpose of using moving average in time series?
12. Which model type uses past values to predict future values?
13. How do we classify data with erratic fluctuations?
14. What is the impact of noise on prediction accuracy?
15. Contrast ARIMA with Naive Forecasting.

Answer Key – Parts 1 to 5 (L1 to L5)

Part 1 – Introduction to Data Analysis & Visualization (L1)

1. **Data** = raw facts, **Information** = processed data, **Insight** = actionable conclusion.
2. Problem definition → Data collection → Cleaning → Analysis → Interpretation → Communication → Decision-making.
3. Descriptive, Diagnostic, Predictive, Prescriptive.
4. Using visual elements to represent data and communicate insights.
5. Descriptive = what happened; Predictive = what will happen.
6. It helps make data-driven decisions based on patterns and trends.
7. Python, R, Excel, Tableau, Power BI.
8. Structured = tables; Unstructured = images, text, audio.
9. Binary: Yes/No, Nominal: City, Gender, Ordinal: Satisfaction scale.

10. Communicating findings using visuals and narrative.
11. Understand patterns, detect outliers, communicate insights.
12. Pandas, Matplotlib, Seaborn.
13. Univariate: one variable; Multivariate: more than one variable.
14. Poor data quality = misleading or incorrect results.
15. Static: snapshot in time; Dynamic: changes over time.

Part 2 – Data Collection & Sampling (L2)

1. Ensures relevant, high-quality data is used in analysis.
2. Population = entire group; Sample = subset used for analysis.
3. Accurately reflects the characteristics of the population.
4. Simple random, stratified, cluster, systematic, convenience, judgmental.
5. Simple random = equal chance; Stratified = divided by strata.
6. Divide into clusters (e.g., cities) and randomly select clusters.
7. To evaluate model performance on unseen data.
8. Choosing a sample that does not represent the whole population.
9. Coverage error, selection bias, non-response, measurement error.
10. When ease is prioritized, but results may be biased.
11. Primary = collected firsthand; Secondary = from existing sources.
12. Stratified sample based on age, gender, region.
13. The list from which a sample is drawn.
14. Quota = fixed number per group; Judgmental = based on expert choice.
15. Some selected participants fail to respond.

Part 3 – Data Presentation (W3)

1. Textual, tabular, graphical.

2. Table showing how often values occur.
3. Qualitative = categories; Quantitative = numeric data.
4. Class = range; Boundaries = ±0.5; Mark = midpoint.
5. (Max – Min) ÷ number of classes.
6. (Frequency ÷ Total) × 100%
7. Bar chart = categories; Histogram = numeric intervals.
8. When displaying part-to-whole relationships (percentages).
9. To show cumulative frequencies.
10. Line graph of class midpoints vs. frequency.
11. Use range and Sturges' rule or equal intervals.
12. Cumulative = running total; Non-cumulative = per class only.
13. Pie charts, bar charts.
14. Detects spread and outliers.
15. Boxplots, histograms help visualize unusual values.

Part 4 – Data Cleaning (W4)

1. Missing data, duplicates, inconsistent formats, noise.
2. Accuracy, completeness, consistency, accessibility.
3. Handle missing data → remove noise → resolve inconsistencies.
4. Sensor failure, unanswered survey items.
5. Drop rows, fill with mean/median/mode, imputation.
6. df['col'].fillna(df['col'].mean())
7. Random errors; handled via binning, smoothing, regression.
8. Binning = grouping; Smoothing = reducing fluctuations.
9. Group similar data points; detect anomalies.
10. Leads to inaccurate counts/statistics.

11. Units mismatch (e.g., inches vs. cm).
12. Drop column or collect data again.
13. df.dropna()
14. Regression models trend to smooth noise.
15. Dirty data misleads models and lowers accuracy.

Part 5 – Data Exploration & Visualization (L5)

1. Analysis of one variable using summary stats.
2. Mean, median, mode.
3. Average of two middle values.
4. Remove top/bottom % to reduce outlier effect.
5. Variance = avg squared diff from mean; Std Dev = sqrt(variance).
6. Q3 – Q1
7. Bell-shaped, symmetric, mean ≈ median ≈ mode.
8. If median ≠ center of box, distribution is skewed.
9. More than one peak in distribution.
10. Correlation = strength of linear relation; Covariance = direction only.
11. Strong positive correlation.
12. Visualize correlations using color-coded matrix.
13. Shows relationships between multiple variables.
14. When comparing frequency across groups.
15. Lossless = shows all data (dotplot); Lossy = summarized (boxplot).

Answer Key – Parts 6 to 8 (L6 to L8)

Part 6 – Data Transformation & Reduction:

1. Min-Max rescales between [0, 1], Z-score standardizes around mean 0 and std 1

2. (30 - 10) / (50 - 10) = 0.5
3. When values vary in scale, e.g., scaling 0–999 → divide by 1000
4. To group numeric data into categorical bins
5. Aggregation combines rows (e.g., daily to monthly), attribute construction creates new columns
6. Parametric uses equations (regression), non-parametric uses grouping/clustering
7. To reduce redundant features and keep variance (PCA)
8. To improve efficiency and avoid overfitting
9. Mean = 0, Standard Deviation = 1
10. Outliers distort the scaling range
11. Normalization = data scaling; compression = data storage reduction
12. Averaging sales per month from daily data
13. Derived attributes give more useful insights
14. Transformation = change structure; reduction = reduce size
15. Parametric reduction

Part 7 – Regression:

1. Predict numeric outcomes from variables
2. Simple = 1 predictor, Multiple = 2+
3. $y = B_o + B_1x + \varepsilon$
4. $(n\Sigma xy – \Sigma x\Sigma y) / (n\Sigma x^2 – (\Sigma x)^2)$
5. As x increases, y decreases
6. 90% variance in y explained by x
7. Average squared prediction error
8. MSE = error, R^2 = % explained
9. Predict house price by size
10. Fit and predict using linear model

11. Coefficient = slope, Intercept = y when x = 0
12. Only after encoding
13. Shows if price increases with quality
14. They distort the regression line
15. No, need non-linear model
16. May overestimate performance
17. Negative slope
18. Skew results, high error
19. Logistic = classification, Linear = prediction
20. Better accuracy and explanation

Part 8 – Time Series:

1. Ordered by time, dependencies matter
2. Constant mean, variance, autocovariance
3. Changing trend, variance, visual plot
4. Trend = long-term direction; Seasonality = repeating pattern
5. Ice cream sales in summer
6. ARMA assumes stationarity
7. AR: past values, MA: past errors
8. ARIMA(p,d,q) = AR, differencing, MA
9. AR=1, diff=1, MA=1
10. Differencing = subtraction; Detrending = remove trend
11. Smoothing data
12. Autoregressive models
13. Irregularity or noise
14. Reduces prediction reliability
15. ARIMA models structure, naive repeats last value