AY 2024-25 Term 1 Mid-Term Checkpoint Quiz

Date / Start Time	1 March 2024/ 1100HRS
Course	**ECON145 Introductory Data Analytics in Healthcare**
Groups	G1
Instructor	Sean Lam
Name of Student	
Student ID	

INSTRUCTIONS TO CANDIDATES

1 The time allowed for this examination paper is **1 Hour**.
2 This is a **CLOSED BOOK** test. **No computers, tablets, smartphones, smart watches, graphical calculators** are allowed in this examination.
3 You are required to answer **ALL** questions.
4 You may use a calculator for performing calculations.

Q1. From the following, which are structured data types found in electronic health records?

i. Patient bed type (e.g., ICU, General Ward).
ii. Doctor notes taken during patient consultation.
iii. Prescribed and dispensed medication dosage information.
iv. Voice recordings of patient histories.

A) i and iii
B) ii and iv
C) iii only
D) i, ii, iv
E) None of the above

ANS: A

Q2. Which of the following sequences most closely represents a correct approach towards data analytics, from the initial stage to the final outcome?

A) Data Collection → Business Question → Analysis Plan → Derive Insights → Recommendations
B) Data Collection → Business Plan → Analysis Plan → Derive Insights → Recommendations
C) Business Plan → Data Collection → Analysis → Recommendations → Derive Insights
D) Data Understanding → Business Question → Data Collection → Data Insights → Recommendations
E) Business Question → Analysis Plan → Data Collection → Derive Insights → Recommendations

ANS: E

Q3. A healthcare organization implements a new electronic medical records (EMR) system focusing on FAIR principles. If they want to ensure the findability of their patient data across various databases, what should they prioritize?

A) Encrypting all data
B) Assigning globally unique and persistent identifiers
C) Decreasing the number of data rows
D) Limiting data access to comply with data governance requirements
E) All of the above

ANS: B

Q4. The healthcare analytics team is asked to identify any correlations between patient characteristics (age, gender, height, number of comorbidities, physical activity level, number of medications, duration on treatment program) and the outcomes of a hypertension treatment program. What type of analysis would be most appropriate for this task?

A) Time-series analysis
B) K-means clustering
C) Multivariate regression
D) Analysis of Variance (ANOVA)
E) None of the above

ANS: C

Q5. In healthcare regression models, which of the following describes the coefficient of determination (R^2)?
A) The difference between the actual and predicted values
B) The correlation coefficient between actual and predicted values
C) The proportion of variance in the dependent variable explained by the independent variables
D) The slope of the regression line
E) The standard deviation of residuals
ANS: C

Q6. In linear regression, what does it mean if a model has multicollinearity?
A) The predictor variables have significant correlations, which distorts the model's accuracy
B) The predictor variables are independent of one another
C) The model predicts multiple outcomes simultaneously
D) The dependent variable has more than one level
E) The residual errors are not normally distributed

ANS: A

Q7. The data science team notices that a large percentage of the hospital readmission records are missing data. Which method would be most appropriate to handle these missing data points in the analysis?
A) Remove all records with missing data
B) Consult the domain experts on the possible reason(s) for these missing data
C) Use imputation techniques to estimate missing values
D) Conduct the analysis without addressing the missing data
E) Replace missing data with zeros

ANS: B

Q8. A healthcare analyst is tasked with predicting patient no-shows for scheduled appointments. What would be the best first step to ensure the predictive model is built on reliable data?
A) Start to create new variables from the data (feature engineering).
B) Perform exploratory data analysis (EDA) to understand the distribution and relationships in the data.
C) Use the data as-is to fit the model comes from a trusted source.
D) Build a simple logistic regression model immediately.
E) Conduct hypothesis testing on the predictions

ANS: B

Q9. An analyst is reviewing the results of a study comparing the effectiveness of three treatment protocols based on the amount of weight loss. What statistical test should the analyst use to determine if there is a significant difference in effectiveness between the two protocols?
A) Chi-square test
B) Paired t-test
C) ANOVA
D) Independent t-test
E) Correlation analysis

ANS: C

Q10. A hospital wants to visualize the relationship between a patient's BMI and the probability of developing Type 2 diabetes. Which of the following visualizations is most appropriate?
A) Bar chart
B) Box plot
C) Scatter plot
D) Line graph
E) Pie chart

ANS: C

Q11. In a healthcare dataset, an analyst notices that patients with higher blood pressure readings tend to have a higher risk of cardiovascular events. What type of variable is the blood pressure reading in this analysis?
A) Dependent variable
B) Categorical variable
C) Independent variable
D) Response variable
E) Outcome variable

ANS: C

Refer to the following R dataframe and answer Q12 – Q20 :

	PatientID	BloodGlucose_2024_01	Cholesterol_2024_01	Height_2024_01	Weight_2024_01_kg	MeanArterialPressure_2024_01
1	ef2d127de3	98 mg/dL	200 mg/dL	5ft7in	72.57 kg	95 mmHg
2	e7f6c01177	115 mg/dL	220 mg/dL	None	102.06 kg	100 mmHg
3	7902699be4	None	195 mg/dL	5ft6in	75.29 kg	92 mmHg
4	2c624232cd	90 mg/dL	185 mg/dL	5ft8in	68.95 kg	89 mmHg
5	19581e27de	120 mg/dL	230 mg/dL	5ft9in	85.73 kg	105 mmHg

	PatientID	BloodGlucose_2024_02	Cholesterol_2024_02	Weight_2024_02_kg
1	ef2d127de3	None	198 mg/dL	73.48 kg
2	e7f6c01177	112 mg/dL	218 mg/dL	101.15 kg
3	7902699be4	103 mg/dL	None	74.38 kg
4	2c624232cd	92 mg/dL	183 mg/dL	68.04 kg
5	19581e27de	118 mg/dL	225 mg/dL	84.82 kg

Q12. In the context of tidy data, why might the current structure of the dataframe be considered untidy?
A) Different variables are combined into each column.
B) The variables for glucose and cholesterol should be combined.
C) The dataframe has unique identifiers for each patient.
D) Many columns have units present (e.g., Mean Arterial Blood Pressure units of mmHg is present in the column).
E) This is a tidy set of data

ANS: A

Q13. What is the most appropriate way to store the data on weight for both time periods to achieve a more tidy dataset?
A) Remove the text in the Weight related columns
B) Create a separate dataframe for the weight data.
C) Reshape the dataframe to have one column for weight, with another column indicating the time period (2024_01 or 2024_2).
D) Average the weight values across both months.
E) Remove weight data for consistency.

ANS: C

Q14. To visualize the relationship between Blood Glucose and Cholesterol levels across patients, which plot would be most appropriate?
A) Scatter plot
B) Box plot
C) Bar chart
D) Line plot
E) Histogram

ANS: A

Q15. Which is the best way to handle missing data in this dataset before performing any analysis, assuming some patients have missing values for Blood Glucose in February?

A) Remove all patients with missing values.

B) Replace missing values with the average Blood Glucose from January.

C) Replace missing values with zero.

D) Perform the analysis with “None” replaced with 0

E) Replace the missing values using imputation techniques based on similar patients' characteristics.

ANS: E

Q16. Which of the following techniques could be used to identify patients who have unusually high or low Mean Arterial Pressure (MAP) across the study period?

A) Calculate the mean MAP across each period and compare.

B) Create a scatter plot with Blood Glucose as the independent variable.

C) Use a box plot to identify outliers in MAP across the study period.

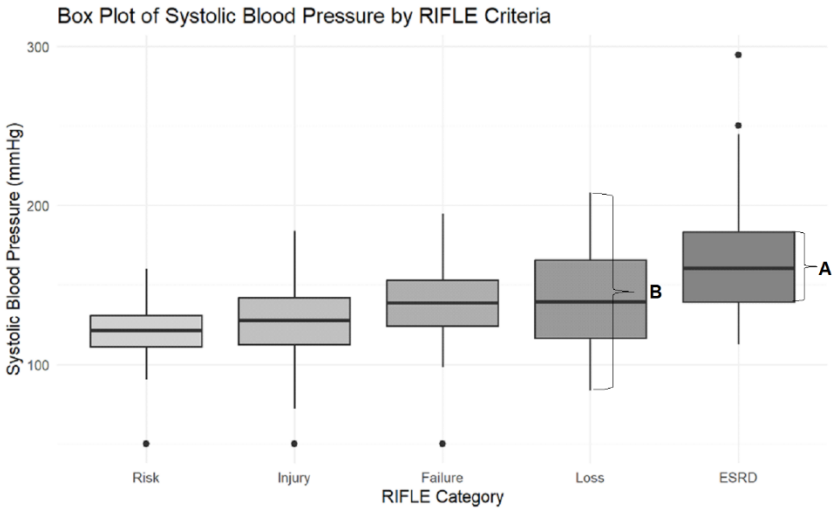
D) Identify all Mean Arterial Pressure values greater than 90 mmHg.

E) Calculate the average Mean Arterial Pressure across both months and remove values that deviate from the mean.

ANS: C

Refer to the following boxplot and answer Q17 – Q19.

The RIFLE criteria is a classification system used to assess the severity of renal failure, and it has 5 levels: Risk, Injury, Failure, Loss, and End-stage renal disease (ESRD).



Q17. What is "A"?

A) The 95% confidence interval
B) The 90% confidence interval
C) The interquartile range (IQR)
D) 3 times the standard deviation
E) 6 time the standard deviation

ANS: C

Q18. What is “B” (representing the span of the whiskers)?

A) Spread of the data within 1.5 times the interquartile range (IQR)
B) Spread of the data within 3 times the IQR
C) Contains observations no more than 1.5 times the IQR from the median
D) Contains observations no more than 1.5 times the IQR from the mean
E) None of the above

ANS: E

Q19. What does a longer whisker in a box plot likely indicate about the data distribution?

A) The presence of more outliers
B) A larger interquartile range (IQR)
C) A greater spread of data or higher variability
D) A symmetric distribution
E) A lower median value

ANS: C

Refer to the following partially completed ANOVA table for Q20-22

	Df	**Sum Sq**	**Mean Sq**	**F value**	**Pr(>F)**
RIFLE	A	48108	B	D	4.48e-10 ***
Residuals	195	167227	C		
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1					

Q20. How many data points are evaluated in this ANOVA test?
A) 4
B) 195
C) 200
D) 250
E) Cannot infer from the table

ANS: 200

Q21. What is the F-Value from this ANOVA – Box D (to 2 d.p.)?
A) 167227.00
B) 12027.00
C) 857.57
D) 14.02
E) 0.07

ANS: D

Q22. Which of the following is the **correct interpretation** of this ANOVA result?
A) There are no significant differences in systolic blood pressure across RIFLE categories.
B) The systolic blood pressure differs across all the RIFLE categories.
C) At least two systolic blood pressure readings differ significantly.
D) The systolic blood pressure is independent of the RIFLE categories.
E) None of the above are correct

ANS. C

Q23. What is the degrees of freedom for RIFLE (Box A)?

A) 3
B) 4
C) 5
D) 200
E) None of the above

ANS. B

Refer to the following R output for Q24 to Q28

```
Call:
lm(formula = Systolic_BP ~ Weight + Glucose + Cholesterol + Smoking + Hematocrit + Vital_Capacity + Age + RIFLE_Categories,
data = data)

Residuals:
Min                 1Q             Median         3Q             Max
-78.541     -20.117         1.004           19.286          101.291

Coefficients:
                    Estimate      Std. Error     t value        Pr(>|t|)
(Intercept)         8.699e+01     1.311e+01      6.634          5.38e-11 ***
Weight              1.914e-01     9.659e-02      1.982          0.0477 *
Glucose             -7.450e-83    6.589e-83      -1.131         0.2585
Cholesterol         2.584e-110    9.128e-110     0.283          0.7772
Smoking             1.791e-01     4.267e-01      -2.060         0.0396 *
Hematocrit          1.917e-01     1.919e-01      0.999          0.3182
Vital_Capacity      6.605e-04     1.064e-03      0.621          0.5347
Age                 -1.051e-01    9.498e-02      -1.107         0.2688
RIFLE_Cat           1.972e-01     9.791e-01      0.201          0.8404
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 29.98 on 991 degrees of freedom
Multiple R-squared:  0.01242, Adjusted R-squared:  0.00445
F-statistic: 1.558 on 8 and 991 DF,  p-value: 0.1332
```

Notes:

1. Systolic_BP: Systolic blood pressure in mmHg
2. Weight: Weight in kg
3. Glucose: Blood glucose in mg/100ml
4. Cholesterol: Cholesterol in mg/100ml
5. Smoking: Number of cigarettes smoked per day
6. Hematocrit: Hematocrit in percentage
7. Vital_Capacity: Vital capacity in centiliters
8. Age: Age in years
9. RIFLE_Cat: RIFLE Categories

Q24. Which of the following variables is statistically significant at the 0.05 level?
A) Systolic BP
B) Smoking
C) Glucose
D) Cholesterol
E) None of the above

Ans: B

Q25. Which predictor has the strongest effect with statistical significance?
A) Weight
B) Glucose
C) Hematocrit
D) RIFLE Categories
E) Insufficient evidence to conclude
Ans: A

Q26. What is the interpretation of the estimate of Weight in the model?
A) For every 1kg increase in weight, the Systolic BP decreases by 1.914 mmHg
B) For every 1kg increase in weight, the Systolic BP increases by 1.914 mmHg
C) For every 10kg increase in weight, the Systolic BP decreases by 1.914 mmHg
D) For every 10kg increase in weight, the Systolic BP increases by 1.914 mmHg
E) All the above are incorrect interpretations

Ans: D

Q27. Which predictor has the strongest evidence suggesting statistical significance?
A) Systolic_BP
B) Age
C) Cholesterol
D) Vital_Capacity
E) Weight

Ans. E

Q28. Considering only Weight, Smoking, Hematocrit, Age and RIFLE categories, what is the estimated systolic BP for a 65-year old patient with a weight of 75kg, a Hematocrit level of 50%, RIFLE category 1 and smoked 5 cigarettes per day?

A) Less than 100
B) 100 to 103
C) More than 103 to 110
D) More than 110 to 120
E) None of the above contains the answer

Ans: C

Q29. A research team is consolidating patient data from several clinical studies. They aim to have a single source of truth that can be used for future studies and research projects. Which of the following aspect of data management is likely the most important for the team to consider?
A) Developing a comprehensive metadata management strategy
B) Establishing a correct statistical analysis plan in their current study protocol
C) Setting up rules and policies to ensure data collection ensure 100% completeness
D) Drafting strict data privacy and security policies in data collection
E) Request for sufficient budget and manpower for a data science team to be set up

Ans: A

Q30. Why are various clinical vocabularies considered complementary in healthcare analytics?
A) Because they reduce the need for data privacy and security
B) Because they facilitate the ease of database design
C) Because they are used together for their intended primary purposes such as clinical inputs and reporting
D) Because they complicate communication between healthcare professionals
E) Because of the need to report the quality of care for patients and service quality for providers

Ans: C

--- End of Paper ---