Data Science Final Exam Review Quiz (60 Questions)

(Answers included at end of document)

Section A: Introduction, Data Science Lifecycle, and Roles (Q1–Q5)

1. **Multiple Choice:** Which of the following phases is **not** one of the main steps in the Data Science Lifecycle process outlined in the course? a) Define Problem b) Acquire Data c) Deploy Model d) Archive Results

2. **Short Answer:** According to the provided definition, what is the primary end goal of Data Science?

3. **Multiple Choice:** Which Python-based library is typically used for Machine Learning modeling? a) Pandas b) NumPy c) Scikit-learn (sklearn) d) Matplotlib

4. **Multiple Choice:** Which role is typically described as an expert in SQL, BI, and Excel, focusing on analyzing data but not necessarily being a domain expert? a) Data Engineer b) Data Scientist c) Data Analyst d) ML Engineer

5. **Multiple Choice:** For a Full-Stack Data Scientist, which of the following falls under Tier 1 – Foundational skills? a) MLOps b) Business acumen c) Data Engineering d) Programming

---

Section B: Python Tools (Q6–Q10)

6. **True/False:** NumPy arrays are generally slower than Python lists for numerical operations.

7. **Definition:** What is a Pandas DataFrame in relation to Series objects?

8. **Short Answer/Code:** What Pandas method replaces missing values with zeros?

9. **Definition:** What is the main purpose of Matplotlib?

10. **Multiple Choice:** The IQR in a boxplot spans which two values? a) Min/Max b) Median/Outlier c) Q1/Q3 d) Mean/Std

---

Section C: Sampling, EDA, and Data Quality (Q11–Q17)

11. Define Exploratory Data Analysis (EDA).

12. Difference between Data Preparation and EDA.
13. Why is Stratified Sampling important for rare events?
14. **Multiple Choice:** What is MCAR? a) Related to observed variables b) Related to missing values c) Missingness occurs purely by chance d) Missingness can be imputed
15. **True/False:** MCAR data can be safely dropped without bias.
16. **Multiple Choice:** If correlation r = 1, what does it mean? a) No relationship b) Strong negative correlation c) Perfect positive linear relationship d) Spurious correlation
17. What |r| values indicate strong correlation?

Section D: Databases and SQL (Q18–Q25)

18. Define DBMS.
19. Name a drawback of using simple file systems instead of a DBMS.
20. What is the purpose of DDL?
21. **Multiple Choice:** Which clause filters aggregated data? a) WHERE b) SELECT c) HAVING d) ORDER BY
22. Which executes first: `FROM` or `WHERE`?
23. **Multiple Choice:** Which join returns only matching rows? a) LEFT b) RIGHT c) FULL OUTER d) INNER
24. What SQL function removes leading/trailing spaces?
25. What does `.fetchall()` return?

Section E: Machine Learning Fundamentals (Q26–Q33)

26. Define Machine Learning regarding learning from data.
27. **Multiple Choice:** Which is unsupervised? a) Predict house prices b) Spam detection c) Topic grouping d) Diabetes classification
28. K-Means is what type of clustering? a) Connectivity b) Density c) Centroid d) Distribution
29. Define WCSS.
30. What indicates optimal K in the Elbow Method?
31. One advantage of Silhouette Score over WCSS.

32. **Multiple Choice:** Which clustering is bottom-up? a) Agglomerative b) Divisive c) DBSCAN d) K-Means++

33. Define a Core Point in DBSCAN.

Section F: Supervised Learning (Q34–Q41)

34. Does Classification predict continuous or categorical values?

35. Why must features be scaled for KNN?

36. **Multiple Choice:** Large k in KNN results in: a) Overfitting b) Underfitting c) Higher training cost d) Switch to Euclidean

37. Two splitting criteria used by Decision Trees.

38. Define a Leaf Node.

39. Logistic Regression maps outputs to what range?

40. Which ensemble method trains models sequentially?

41. What are OOB points used for in Random Forests?

Section G: Evaluation Metrics (Q42–Q49)

42. Define Bias.

43. What is the goal of the bias-variance tradeoff?

44. What is the purpose of the Test dataset?

45. Advantage of RMSE over MSE.

46. **Multiple Choice:** Accuracy numerator includes: a) TP + FP b) TP + TN c) TN + FN d) Only TP

47. **Multiple Choice:** For medical diagnostics, prioritize: a) Accuracy b) Precision c) Recall d) F1

48. What two rates define the ROC curve axes?

49. Meaning of AUC = 1.0?

Section H: Artificial Neural Networks (Q50–Q52)

50. In $y = f(\sum w_i x_i + b)$, what is b?

51. Purpose of the Activation Function.

52. **Multiple Choice:** ReLU is known for: a) Squashing to 0–1 b) Being linear c) Fast, sparse, avoids vanishing gradients d) Computationally costly

---

Section I: Generative AI and Ethics (Q53–Q60)

53. Primary goal of Generative AI.
54. What core function do LLMs perform instead of “thinking”?
55. Define Retrieval-Augmented Generation.
56. What is an AI hallucination?
57. **Multiple Choice:** “As a financial analyst…” is what technique? a) Few-Shot b) Zero-Shot c) Role-Playing d) Iterative Refinement
58. Underlying architecture of GPT.
59. Compare Initial Training vs. Fine-Tuning.
60. Difference between ANI and AGI.

Answer Key

1. d
2. Extracting knowledge/actionable insights
3. c
4. c
5. d
6. False
7. A 2D structure composed of aligned Series
8. `fillna(0)`
9. Creating visualizations
10. c
11. Data exploration to summarize characteristics
12. Prep = fixing data; EDA = exploring data
13. Ensures representation of rare events

14. c

15. True

16. c

17. 0.7–1.0

18. Software that reads/writes data and ensures integrity

19. Redundancy, inconsistency, poor access, concurrency issues

20. Define schema (e.g., CREATE TABLE)

21. c

22. FROM

23. d

24. `TRIM()`

25. List of tuples

26. Improves performance with more data

27. c

28. c

29. Measures intra-cluster compactness

30. The "elbow" point

31. Measures both compactness and separation

32. a

33. Point with ≥ MinPts within radius Eps

34. Categorical

35. Distance-based algorithm sensitive to scale

36. b

37. Gini Impurity; Entropy

38. Final decision node

39. 0–1

40. Boosting

41. Internal validation set

42. Error from oversimplification

43. Optimal complexity balancing bias & variance

44. Final model evaluation

45. Same units as target variable

46. b

47. c (Recall)

48. TPR vs. FPR

49. Perfect classifier

50. Bias term

51. Introduces non-linearity

52. c

53. Generate realistic, relevant content

54. Predict next word

55. RAG

56. Hallucination

57. c

58. Transformer with attention

59. Initial Training = general patterns; Fine-Tuning = task-specific

60. ANI = single task; AGI = human-like general intelligence