Data Analysis - Exam Questions

Objective Questions (MCQ/MSQ) - 20 Questions

1. Which Excel function calculates the correlation coefficient between two ranges?

A) `CORRELATION()`
B) `CORREL()`
C) `PEARSON()`
D) `RELATE()`

Answer: B

2. In DuckDB, which clause is used for window functions?

A) `WINDOW`
B) `OVER`
C) `PARTITION`
D) `FRAME`

Answer: B

3. Which Python function tests for statistical significance in correlation?

A) `scipy.stats.pearsonr()`
B) `pandas.corr()`
C) `numpy.corrcoef()`
D) `statsmodels.correlation()`

Answer: A

4. What does the IQR method use to detect outliers?

A) Mean and standard deviation
B) Q1 and Q3 quartiles
C) Median and mode
D) Min and max values

Answer: B

5. Which Excel function performs multiple linear regression?

A) `REGRESSION()`
B) `LINEST()`
C) `TREND()`
D) `FORECAST()`

Answer: B

6. In geospatial analysis, which Python library calculates distances between coordinates?

A) `geopandas`
B) `folium`
C) `geopy`
D) `shapely`

Answer: C

7. Which NetworkX function calculates betweenness centrality?

A) `nx.centrality()`
B) `nx.betweenness_centrality()`
C) `nx.degree_centrality()`
D) `nx.closeness_centrality()`

Answer: B

8. What is the primary advantage of Parquet over CSV for analytics?

A) Human readable format
B) Columnar storage and compression
C) Easier to edit
D) Better for small files

Answer: B

9. Which Excel quartile function is recommended for outlier detection?

A) `QUARTILE()`
B) `QUARTILE.INC()`
C) `QUARTILE.EXC()`
D) `PERCENTILE()`

Answer: B

10. In pandas, which method creates pivot tables?

A) `df.pivot()`
B) `df.pivot_table()`
C) `df.crosstab()`
D) `df.groupby()`

Answer: B

11. Which QGIS file format stores geographic boundaries?

A) KML
B) Shapefile
C) GeoJSON
D) All of the above

Answer: D

12. What does R-squared measure in regression analysis?

A) Correlation strength
B) Variance explained by the model
C) Statistical significance
D) Prediction accuracy

Answer: B

13. Which DuckDB function reads Parquet files directly?

A) `read_parquet()`
B) `FROM parquet_scan()`
C) `SELECT * FROM 'file.parquet'`
D) All of the above

Answer: D

14. In network analysis, what does the Louvain algorithm detect?

A) Shortest paths
B) Communities/clusters
C) Central nodes
D) Network diameter

Answer: B

15. Which Excel function forecasts future values using exponential smoothing?

A) `FORECAST()`
B) `FORECAST.ETS()`
C) `TREND()`
D) `GROWTH()`

Answer: B

16. What is the correct formula for coefficient of variation?

A) `Standard deviation / Mean`
B) `Mean / Standard deviation`
C) `Variance / Mean`
D) `Mean / Variance`

Answer: A

17. Which Python library is best for interactive geospatial visualization?

A) `matplotlib`
B) `seaborn`
C) `folium`
D) `plotly`

Answer: C

18. In SQL window functions, what does `ROWS BETWEEN 6 PRECEDING AND CURRENT ROW` create?

A) 6-row partition
B) 7-row moving window
C) 6-row lag
D) Current row only

Answer: B

19. Which method handles multiple testing correction?

A) Bonferroni correction
B) Benjamini-Hochberg FDR
C) Both A and B
D) Neither A nor B

Answer: C

20. What does `geodesic` distance measure?

A) Straight-line distance
B) Distance on Earth's surface
C) Manhattan distance
D) Euclidean distance

Answer: B

Subjective/Scenario Questions - 20 Questions

1. Statistical Analysis Design

You're analyzing customer purchase data to identify factors affecting sales. The dataset has 50,000 records with potential outliers and missing values. Design a comprehensive analysis approach including outlier detection, correlation analysis, and regression modeling.

Answer: (1) Use IQR method for outlier detection with `QUARTILE.INC()` in Excel or `scipy.stats` in Python, (2) Handle missing values with appropriate imputation or exclusion, (3) Calculate correlation matrix to identify multicollinearity, (4) Perform multiple regression with `LINEST()` or `sklearn`, (5) Validate assumptions through residual analysis, (6) Apply multiple testing correction for significance tests.

2. Tool Selection for Large Dataset Analysis

Compare DuckDB, pandas, and Excel for analyzing a 10GB sales dataset. Consider performance, memory usage, and analytical capabilities. Which would you choose for different scenarios?

Answer: DuckDB: Best for large datasets, SQL familiarity, columnar processing. **Pandas**: Good for medium datasets, flexible transformations, rich ecosystem. **Excel**: Limited to small datasets but excellent for quick exploration and presentation. Choose DuckDB for 10GB+ data, pandas for complex transformations on smaller subsets, Excel for final reporting and stakeholder communication.

3. Geospatial Analysis Strategy

Design an analysis to determine optimal locations for new retail stores based on competitor locations, population density, and transportation access. What tools and methods would you use?

Answer: Use **QGIS** for spatial data management and buffer analysis, **Python GeoPandas** for programmatic analysis, **Folium** for interactive visualization. Methods: (1) Buffer analysis around competitors, (2) Population density overlay, (3) Transportation network analysis, (4) Site scoring algorithm combining factors, (5) Sensitivity analysis for different weights.

4. Network Analysis Application

Analyze collaboration patterns in a research organization using publication co-authorship data. Design your approach to identify key researchers, research clusters, and collaboration opportunities.

Answer: (1) Build co-authorship network using **NetworkX**, (2) Calculate centrality measures (betweenness, closeness, degree), (3) Apply community detection (Louvain algorithm), (4) Identify bridge nodes connecting clusters, (5) Visualize with **Kumu** or **Gephi**, (6) Recommend collaboration strategies based on network gaps.

5. Regression Analysis Validation

You've built a regression model predicting house prices with $R^2 = 0.85$ and $p < 0.001$. What additional validation steps would you perform to ensure model reliability?

Answer: (1) Check residual plots for homoscedasticity and normality, (2) Test for multicollinearity using VIF, (3) Validate assumptions through diagnostic plots, (4) Perform cross-validation, (5) Test on holdout dataset, (6) Check for influential outliers using Cook's distance, (7) Assess practical significance vs. statistical significance.

6. Time Series Forecasting

Design a forecasting system for monthly sales data with seasonal patterns and trend components. Compare different approaches and recommend the best strategy.

Answer: Compare **Excel FORECAST.ETS()** for simplicity, **Python statsmodels** for advanced methods, and **Prophet** for automatic seasonality detection. Recommend: (1) Decompose series into trend/seasonal/residual, (2) Use ETS or ARIMA for statistical approach, (3) Validate with cross-validation, (4) Combine multiple models for ensemble forecasting, (5) Monitor forecast accuracy continuously.

7. Data Quality Assessment

Create a comprehensive data quality framework for analytical datasets. What metrics would you track and how would you implement automated quality checks?

Answer: Track: (1) **Completeness** (null rates), (2) **Validity** (format compliance), (3) **Consistency** (cross-field validation), (4) **Accuracy** (business rule compliance), (5) **Timeliness** (data freshness). Implement using **Great Expectations** or custom **pandas** functions with automated reporting and threshold-based alerting.

8. Performance Optimization

Your correlation analysis on a 1M row dataset takes 30 minutes in pandas. How would you optimize this for better performance?

Answer: (1) Use **DuckDB** for columnar processing, (2) Sample data for exploratory analysis, (3) Use **Dask** for parallel processing, (4) Optimize data types and memory usage, (5) Cache intermediate results, (6) Use **NumPy** operations where possible, (7) Consider **Polars** as pandas alternative.

9. Statistical Significance Testing

Design a testing framework for A/B experiments with multiple metrics. How would you handle multiple comparisons and ensure statistical rigor?

Answer: (1) Pre-register hypotheses and analysis plan, (2) Apply **Bonferroni** or **Benjamini-Hochberg** correction for multiple testing, (3) Use appropriate statistical tests (t-test, chi-square, Mann-Whitney), (4) Calculate effect sizes alongside p-

values, (5) Implement sequential testing for early stopping, (6) Report confidence intervals and practical significance.

10. Geospatial Data Integration

Combine demographic data, geographic boundaries, and point-of-interest data for market analysis. What challenges would you face and how would you address them?

Answer: Challenges: (1) **Coordinate system mismatches** - standardize to common CRS, (2) **Scale differences** - use appropriate aggregation levels, (3) **Data quality** - validate coordinates and boundaries, (4) **Performance** - use spatial indexing and efficient joins. Solutions: Use **GeoPandas** for data integration, **QGIS** for validation, and **PostGIS** for large-scale processing.

11. Outlier Treatment Strategy

In a financial dataset, you've identified outliers that could be either data errors or legitimate extreme values. Design a systematic approach to handle them.

Answer: (1) **Investigate context** - domain knowledge and business rules, (2) **Multiple detection methods** - IQR, Z-score, isolation forest, (3) **Validation** - cross-reference with external sources, (4) **Treatment options** - removal, transformation, or robust methods, (5) **Impact analysis** - compare results with/without outliers, (6) **Documentation** - record decisions and rationale.

12. Cross-Platform Analytics

Design an analytics workflow that works across Excel, Python, and SQL environments for different stakeholders. How would you ensure consistency and reproducibility?

Answer: (1) **Standardized data formats** - use Parquet for interchange, (2) **Common calculations** - document formulas and logic, (3) **Version control** - Git for code, data versioning for datasets, (4) **Automated testing** - validate results across platforms, (5) **Documentation** - clear specifications and examples, (6) **Training** - ensure team understands each platform's strengths.

13. Real-time Analytics

Adapt your batch correlation analysis to work with streaming data. What architectural changes would you make?

Answer: (1) **Streaming framework** - Apache Kafka + Spark Streaming, (2) **Windowed calculations** - sliding/tumbling windows for correlations, (3) **Incremental updates** - update statistics without full recalculation, (4) **State management** - maintain running statistics, (5) **Alerting** - real-time notifications for significant changes, (6) **Visualization** - live dashboards with streaming updates.

14. Model Interpretability

Your regression model has high accuracy but stakeholders need to understand which factors drive the predictions. How would you enhance interpretability?

Answer: (1) **Feature importance** - calculate coefficient magnitudes and p-values, (2) **Partial dependence plots** - show individual feature effects, (3) **SHAP values** - explain individual predictions, (4) **Simplified models** - create interpretable approximations, (5) **Visualization** - clear charts showing relationships, (6) **Documentation** - plain language explanations of findings.

15. Data Governance

Implement data governance for analytical datasets including lineage tracking, access controls, and audit trails. What framework would you establish?

Answer: (1) **Data catalog** - metadata management and discovery, (2) **Lineage tracking** - document data flow and transformations, (3) **Access controls** - role-based permissions and authentication, (4) **Quality monitoring** - automated data quality checks, (5) **Audit logging** - track data access and modifications, (6) **Compliance** - ensure regulatory requirements are met.

16. Collaborative Analysis

Design a collaborative analytics environment where multiple analysts can work on the same datasets without conflicts. What tools and processes would you

implement?

Answer: (1) **Version control** - Git for code, DVC for data, (2) **Shared environments** - JupyterHub or cloud notebooks, (3) **Data standards** - common schemas and naming conventions, (4) **Communication** - documentation and code reviews, (5) **Resource management** - shared compute and storage, (6) **Conflict resolution** - merge strategies and testing protocols.

17. Scalability Planning

Your current analysis runs on a single machine but data volume is growing 10x annually. Design a scalable architecture for the next 3 years.

Answer: (1) **Distributed computing** - migrate to Spark or Dask, (2) **Cloud infrastructure** - auto-scaling compute resources, (3) **Data partitioning** - optimize for query patterns, (4) **Caching strategies** - reduce redundant computations, (5) **Monitoring** - track performance and costs, (6) **Gradual migration** - phase transition to minimize disruption.

18. Automated Insights

Create an automated system that identifies interesting patterns and anomalies in business data. What approach would you take?

Answer: (1) **Statistical monitoring** - track key metrics and detect changes, (2) **Anomaly detection** - isolation forest, statistical process control, (3) **Pattern recognition** - clustering and association rules, (4) **Natural language generation** - automated insight summaries, (5) **Alerting system** - prioritized notifications, (6) **Feedback loop** - learn from user interactions.

19. Regulatory Compliance

Ensure your analytics comply with data protection regulations (GDPR, CCPA). What measures would you implement?

Answer: (1) **Data minimization** - collect only necessary data, (2) **Anonymization** - remove or mask PII, (3) **Consent management** - track data usage permissions, (4) **Right to deletion** - implement data removal processes, (5) **Audit trails** - log all

data processing activities, (6) **Privacy by design** - build compliance into workflows.

20. Business Impact Measurement

Design a framework to measure and communicate the business impact of your analytics initiatives. What metrics and reporting would you establish?

Answer: (1) **KPI alignment** - link analytics to business objectives, (2) **ROI calculation** - quantify cost savings and revenue impact, (3) **Decision tracking** - monitor how insights influence actions, (4) **A/B testing** - measure improvement from analytics-driven changes, (5) **Stakeholder feedback** - regular surveys and interviews, (6) **Success stories** - document and share impactful use cases.