Exam 1: SAS Big Data Preparation, Statistics, and Visual Exploration

Data Management - 50%

Navigate within the Data Management Studio Interface

- Register a new QKB
- Create and connect to a repository
- Define a data connection
- Specify Data Management Studio options
- Access the QKB
- Create a name value macro pair
- Access the business rules manager
- Access the appropriate monitoring report
- Attach and detach primary tabs

Create, design and be able to explore data explorations and interpret results

Define and create data collections from exploration results

Create and explore a data profile

- Create a data profile from different sources (text file, filtered table, SQL query)
- Interpret results (frequency distribution & pattern)
- Use collections from profile results

Design data standardization schemes

- Build a scheme from profile results
- Build a scheme manually
- Update existing schemes

Create Data Jobs

- Rename output fields
- Add nodes and preview nodes
- Run a data job
- View a log and settings
- Work with data job settings and data job displays
Exam Content Guide

- Best practices (how do you ensure that you are following a particular best practice): examples: insert notes, establish naming conventions
- Work with branching
- Join tables
- Apply the Field layout node to control field order
- Work with the Data Validation node:
  - Add it to the job flow
  - Specify properties/review properties
  - Edit settings for the Data Validation node
- Work with data inputs
- Work with data outputs
- Profile data from within data jobs
- Interact with the Repository from within Data Jobs
- Determine how data is processed
- Data job variables
- Set Sorting properties for the Data Sorting node
  - Set appropriate advanced properties options for the Data Sorting Node

Apply a Standardization definition and scheme

- Use a definition
- Use a scheme
- Be able to determine the differences between definition and scheme
- Explain what happens when you use both a definition and scheme
- Review and interpret standardization results
- Be able to explain the different steps involved in the process of standardization

Apply Parsing definitions

- Distinguish between different data types and their tokens
- Review and interpret parsing results
- Be able to explain the different steps involved in the process of parsing
- Use parsing definition

Compare and contrast the differences between identification analysis and right fielding nodes

- Review results
- Explain the technique used for identification (process of the definition)

Apply the Gender Analysis node to determine gender

- Use gender definition
- Interpret results
- Explain different techniques for accomplishing gender analysis
Create an Entity Resolution Job

- Use a node in the data job that is the clustering node and explain why you would want to use it
- Survivorship (surviving record identification)
  - Record rules
  - Field rules
  - Options for survivorship
- Discuss and apply the Cluster Diff node
- Apply Cross-field matching (new option)
- Use the Match Codes Node to select match definitions for selected fields
  - Outline the various uses for match codes (join)
  - Use the definition
  - Interpret the results
  - Match versus match parsed
  - Explain the process for creating a match code
  - Select sensitivity for a selected match definition
  - Apply matching best practices

Define and create business rules

- Use Business Rules Manager
- Create a new business rule
  - Name/label rule
  - Specify type of rule
  - Define checks
  - Specify fields
- Distinguish between different types of business rules
  - Row
  - Set
  - Group
- Apply business rules
  - Profile
  - Execute business rule node
- Use of Expression Builder
- Apply best practices

Describe the organization, structure and basic navigation of the QKB

- Identify and describe locale levels (global, language, country)
- Navigate the QKB (tab structure, copy definitions, etc.)
- Identify data types and tokens
Be able to articulate when to use the various components of the QKB

- Components include:
  - Regular expressions
  - Schemes
  - Phonetics library
  - Vocabularies
  - Grammar
  - Chop Tables

Define the processing steps and components used in the different definition types

- Identify/describe the different definition types
  - Parsing
  - Standardization
  - Match
  - Identification
  - Casing
  - Extraction
  - Locale guess
  - Gender
  - Patterns
ANOVA and Regression - 30%

Verify the assumptions of ANOVA

- Explain the central limit theorem and when it must be applied
- Examine the distribution of continuous variables (histogram, box-whisker, Q-Q plots)
- Describe the effect of skewness on the normal distribution
- Define H0, H1, Type I/II error, statistical power, p-value
- Describe the effect of sample size on p-value and power
- Interpret the results of hypothesis testing
- Interpret histograms and normal probability charts
- Draw conclusions about your data from histogram, box-whisker, and Q-Q plots
- Identify the kinds of problems may be present in the data: (biased sample, outliers, extreme values)
- For a given experiment, verify that the observations are independent
- For a given experiment, verify the errors are normally distributed
- Use the UNIVARIATE procedure to examine residuals
- For a given experiment, verify all groups have equal response variance
- Use the HOVTEST option of MEANS statement in PROC GLM to assess response variance

Analyze differences between population means using the GLM and TTEST procedures

- Use the GLM Procedure to perform ANOVA
  - CLASS statement
  - MODEL statement
  - MEANS statement
  - OUTPUT statement
- Evaluate the null hypothesis using the output of the GLM procedure
- Interpret the statistical output of the GLM procedure (variance derived from MSE, F value, p-value R², Levene's test)
- Interpret the graphical output of the GLM procedure
- Use the TTEST Procedure to compare means

Perform ANOVA post hoc test to evaluate treatment affect

- use the LSMEANS statement in the GLM or PLM procedure to perform pairwise comparisons
- use PDIFF option of LSMEANS statement
- use ADJUST option of the LSMEANS statement (TUKEY and DUNNETT)
- Interpret diffograms to evaluate pairwise comparisons
Interpret control plots to evaluate pairwise comparisons

Compare/Contrast use of pairwise T-Tests, Tukey and Dunnett comparison methods

PLM

Detect and analyze interactions between factors

Use the GLM procedure to produce reports that will help determine the significance of the interaction between factors.
  - MODEL statement
  - LSMEANS with SLICE=option (Also using PROC PLM)
  - ODS SELECT

Interpret the output of the GLM procedure to identify interaction between factors:
  - p-value
  - F Value
  - R Squared
  - TYPE I SS
  - TYPE III SS

Fit a multiple linear regression model using the REG and GLM procedures

Use the REG procedure to fit a multiple linear regression model

Use the GLM procedure to fit a multiple linear regression model

Analyze the output of the REG, PLM, and GLM procedures for multiple linear regression models

Interpret REG or GLM procedure output for a multiple linear regression model:
  - convert models to algebraic expressions
  - Convert models to algebraic expressions
  - Identify missing degrees of freedom
  - Identify variance due to model/error, and total variance
  - Calculate a missing F value
  - Identify variable with largest impact to model
  - For output from two models, identify which model is better
  - Identify how much of the variation in the dependent variable is explained by the model
  - Conclusions that can be drawn from REG, GLM, or PLM output: (about H0, model quality, graphics)

Use the REG or GLMSELECT procedure to perform model selection

Use the SELECTION option of the model statement in the GLMSELECT procedure

Compare the different model selection methods (STEPWISE, FORWARD, BACKWARD)

Enable ODS graphics to display graphs from the REG or GLMSELECT procedure

Identify best models by examining the graphical output (fit criterion from the REG or GLMSELECT procedure)
Exam Content Guide

- Assign names to models in the REG procedure (multiple model statements)

Assess the validity of a given regression model through the use of diagnostic and residual analysis

- Explain the assumptions for linear regression
- From a set of residuals plots, assess which assumption about the error terms has been violated
- Use REG procedure MODEL statement options to identify influential observations (Student Residuals, Cook's D, DFFITS, DFBETAS)
- Explain options for handling influential observations
- Identify collinearity problems by examining REG procedure output
- Use MODEL statement options to diagnose collinearity problems (VIF, COLLIN, COLLINOINT)

Perform logistic regression with the LOGISTIC procedure

- Identify experiments that require analysis via logistic regression
- Identify logistic regression assumptions
- Logistic regression concepts (log odds, logit transformation, sigmoidal relationship between p and X)
- Use the LOGISTIC procedure to fit a binary logistic regression model (MODEL and CLASS statements)

Optimize model performance through input selection

- Use the LOGISTIC procedure to fit a multiple logistic regression model
- LOGISTIC procedure SELECTION=SCORE option
- Perform Model Selection (STEPWISE, FORWARD, BACKWARD) within the LOGISTIC procedure

Interpret the output of the LOGISTIC procedure

- Interpret the output from the LOGISTIC procedure for binary logistic regression models:
  - Model Convergence section
  - Testing Global Null Hypothesis table
  - Type 3 Analysis of Effects table
  - Analysis of Maximum Likelihood Estimates table
  - Association of Predicted Probabilities and Observed Responses
Visual Data Exploration - 20%

Examine, modify, and create data items

- Create and use parameterized data items
- Examine data item properties and measure details
- Change data item properties
- Create custom sorts
- Create distinct counts
- Create aggregated measures
- Create calculated items
- Create hierarchies
- Create custom categories

Select and work with data sources

- Work with multiple data sources
- Change data sources
- Refresh data sources

Create, modify, and interpret automatic chart visualizations in Visual Analytics Explorer

- Identify default visualizations
- Identify the properties available in an automatic chart

Create, modify, and interpret graph and table visualizations in Visual Analytics Explorer

- Work with list table visualizations
- Work with crosstab visualizations
- Work with bar chart visualizations
- Work with line chart visualizations
- Work with scatter plot visualizations
- Work with bubble plot visualizations
- Work with histogram visualizations
- Work with box plot visualizations
- Work with heat map visualizations
- Work with geo map visualizations
- Work with treemap visualizations
- Work with correlation matrix visualizations
Enhance visualizations with analytics within Visual Analytics Explorer

- Add fit lines to visualizations
- Create forecasts
- Interpret word clouds

Interact with visualizations and explorations within Visual Analytics Explorer

- Control appearance of visualizations within explorations
- Add comments to visualizations and explorations
- Use filters on data source and visualizations
- Share explorations
- Share visualizations

Note: All 32 main objectives will be tested on every exam. The 210 expanded objectives are provided for additional explanation and define the entire domain that could be tested.