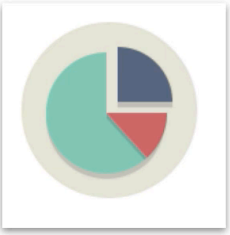
Whizlabs - ML Specialty Exam Course - Data Analysis - part 3

Analyze and visualize Data for Machine Learning

AWS Machine Learning Exploratory Data Analysis

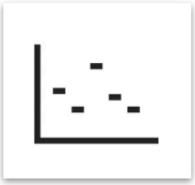
Analyzing and Visualizing Data for Machine Learning

- Visualize data before choosing a machine learning algorithm
  - Identify patterns
  - Find corrupt data
  - Identify outliers
  - Find imbalances in the data
  - Explore and demonstrate important relationships, and strength of relationships (density), using plots/charts



Charting Data for Machine Learning

- Types of information to convey via Business Intelligence (BI) tools
  - Key Performance Indicators (KPIs)
  - Relationships
  - Comparisons
  - Distributions
  - Compositions



Charting Data - Key Performance Indicators (KPIs)

- A single value that represents a particular area or function and shows relative performance
  - Net Promoter Score (NPS)
  - Customer Profitability Score (CPS)
  - Conversion Rate
  - Relative Market Share
  - Net Profit Margin

- Use KPI charts to represent these indicators

Charting Data - Relationships

- Establish or prove a relationship between 2 or more variables
- Best chart to use depends on number of variables being compared
  - Scatter Chart
    - Two variables, example: social media spend to adoption rate

  - Bubble Chart
    - Three variables, example: comparing investment return, investment duration, and investment commitment

Charting Data - Comparisons

- Show how variables change over time or show a static view of how different variables compare
- Best chart to use depends on number of variables being compared

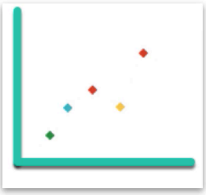
  - Bar Chart
    - One variable, example: website hits in a given month
  - Table
    - Three variables, example: two dimensions represented as the columns and rows, the third by the data in the cells
  - Column Chart

    - One or two variables changing over time, example: show year-over-year sales and number of marketing campaigns
  - Line Chart
    - Three or more variables changing over time, example: show year-over-year sales, number of marketing campaigns, and web traffic

Charting Data - Distributions

- Show how data is distributed over defined intervals, Interval meaning clustering or grouping, not time
  - Column Histogram

    - One variable, example: showing how many voters are in various generation groups
    - Counting something and putting them into buckets
  - Scatter Chart



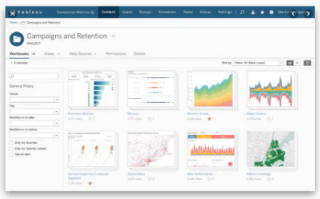
    - Two variables, example: relating return on investment, investment duration, and investment size
    - X-axis is investment time, y-axis is return on investment, and the bubble size is the investment size

Charting Data - Compositions

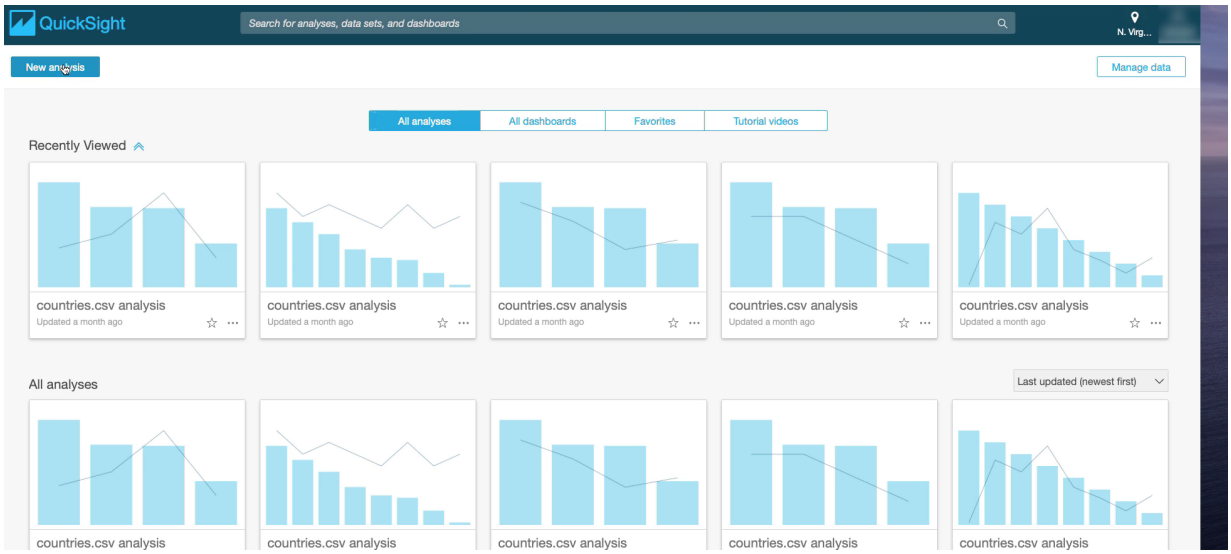
- Show the elements that make up data set, static or changing over time
  - Pie Chart: simple share of total
  - Stacked 100% Bar Chart: components of components
  - Tree Map: share of total
  - Stacked Area Chart: 5 or more periods; relative and absolute differences
  - Stacked 100% Area Chart: 5 or more periods; relative differences
  - Stacked Column Chart: less than 5 periods; relative and absolute differences
  - Stacked 100% Column Chart: less than 5 periods; relative differences

Business Intelligence Tools

- Amazon Quicksight
  - Cloud powered business intelligence service
  - Create interactive dashboards that include machine learning insights
    - Anomaly detection, forecasting, auto-narratives
- TensorFlow with TensorBoard
- Tableau



Data Visualization - Lab



Use a Manifest file to describe where is the data in S3

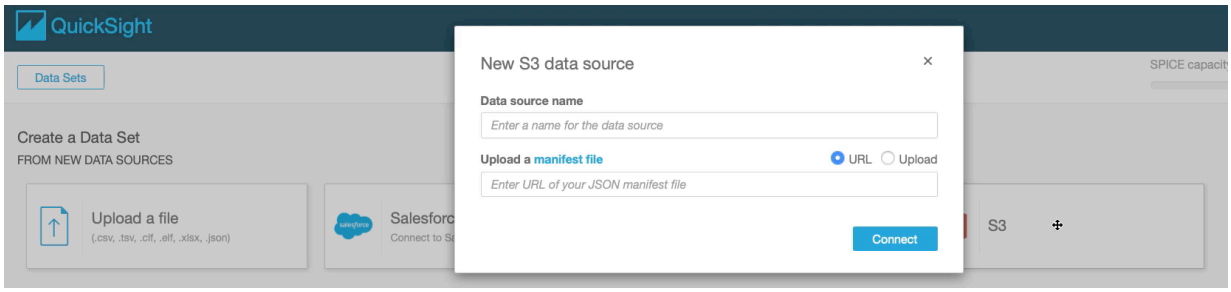
```
{
  "fileLocations": [
    {
      "URIs": [
        "s3://machine-learning-exam/FAO-database.csv"
      ]
    }
  ],
  "globalUploadSettings": {
    "format": "CSV",
    "delimiter": ",",
    "textqualifier": "'",
    "containsHeader": "true"
  }
}
```

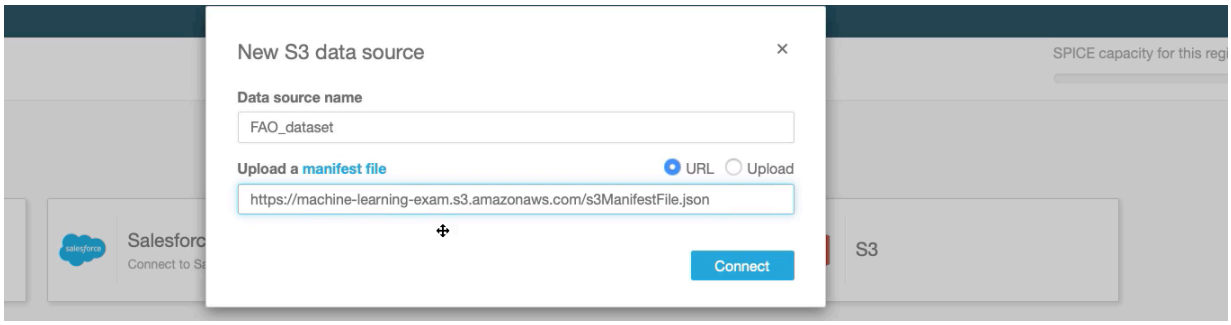
Store that manifest file in S3

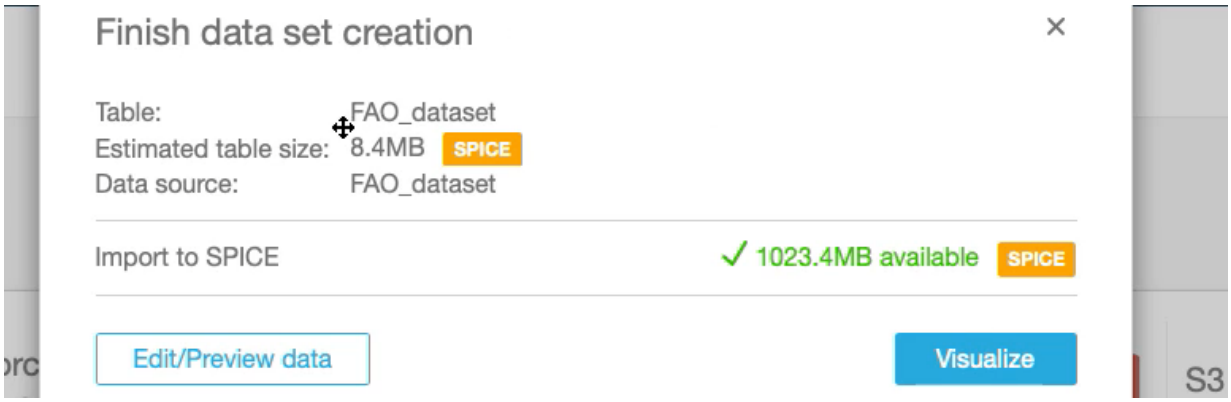
Object URL

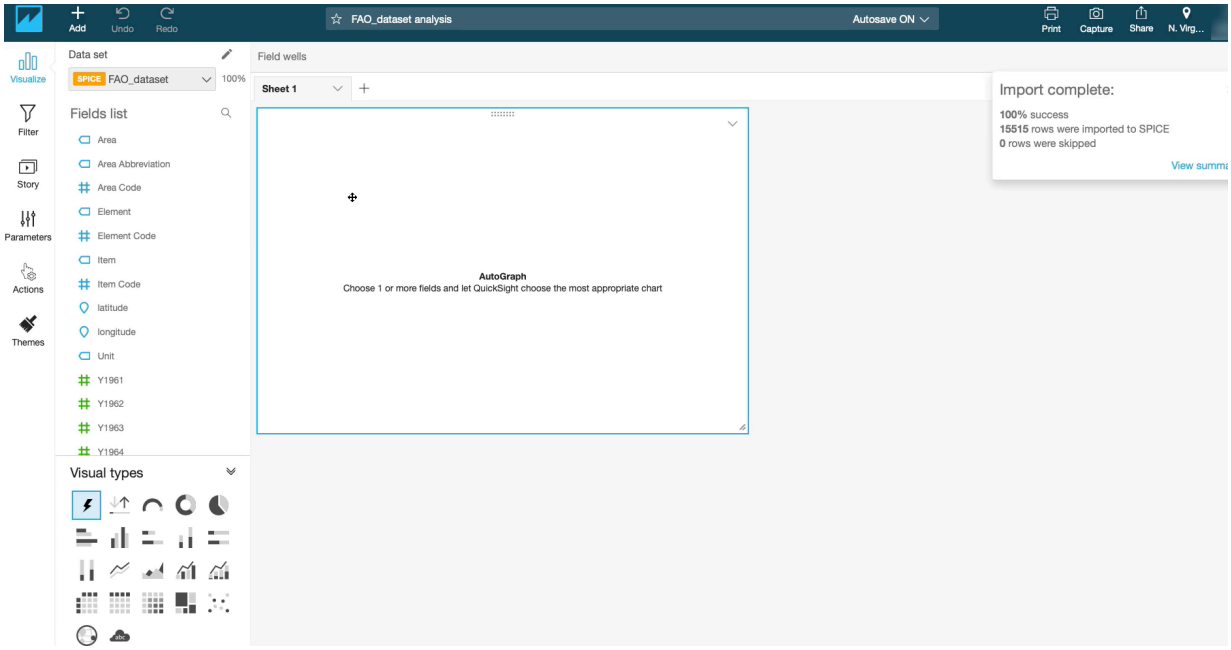
https://machine-learning-exam.s3.amazonaws.com/s3ManifestFile.json

Create new dataset. From S3

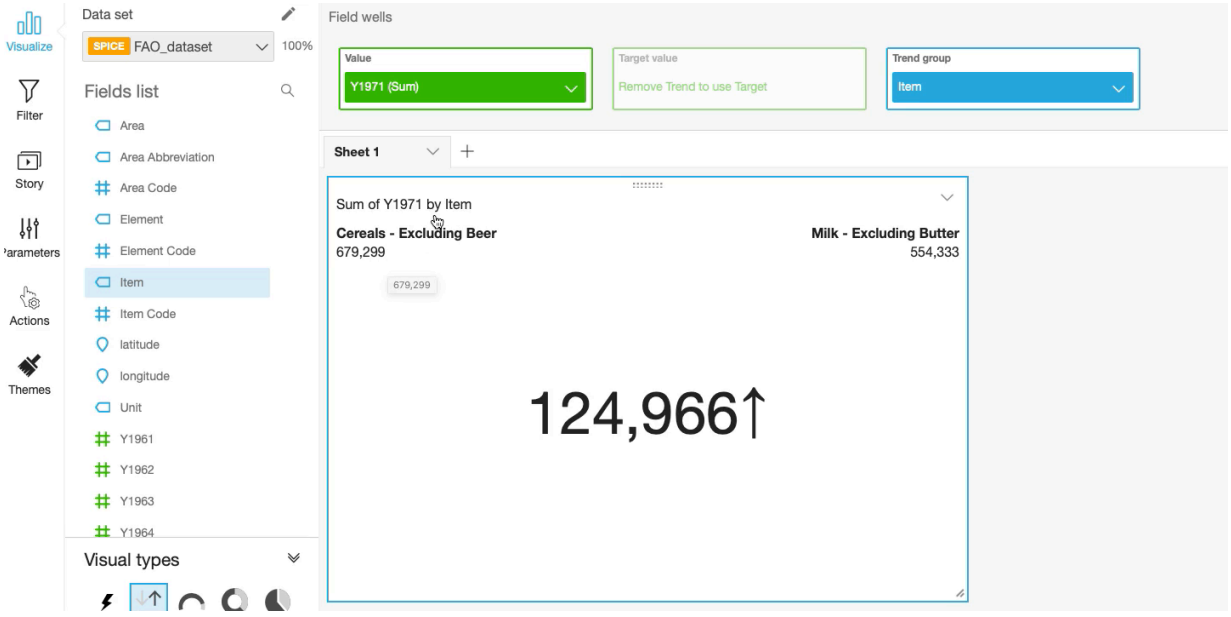

QuickSight
Data Sets
Create a Data Set
FROM NEW DATA SOURCES
Upload a file
(.csv, .tsv, .clf, .elf, .xlsx, .json)
New S3 data source
Data source name
Enter a name for the data source
Upload a manifest file
URL
Upload
Enter URL of your JSON manifest file
Connect
S3


New S3 data source
Data source name
FAO_dataset
Upload a manifest file
URL
Upload
https://machine-learning-exam.s3.amazonaws.com/s3ManifestFile.json
Connect
SPICE capacity for this regi
S3

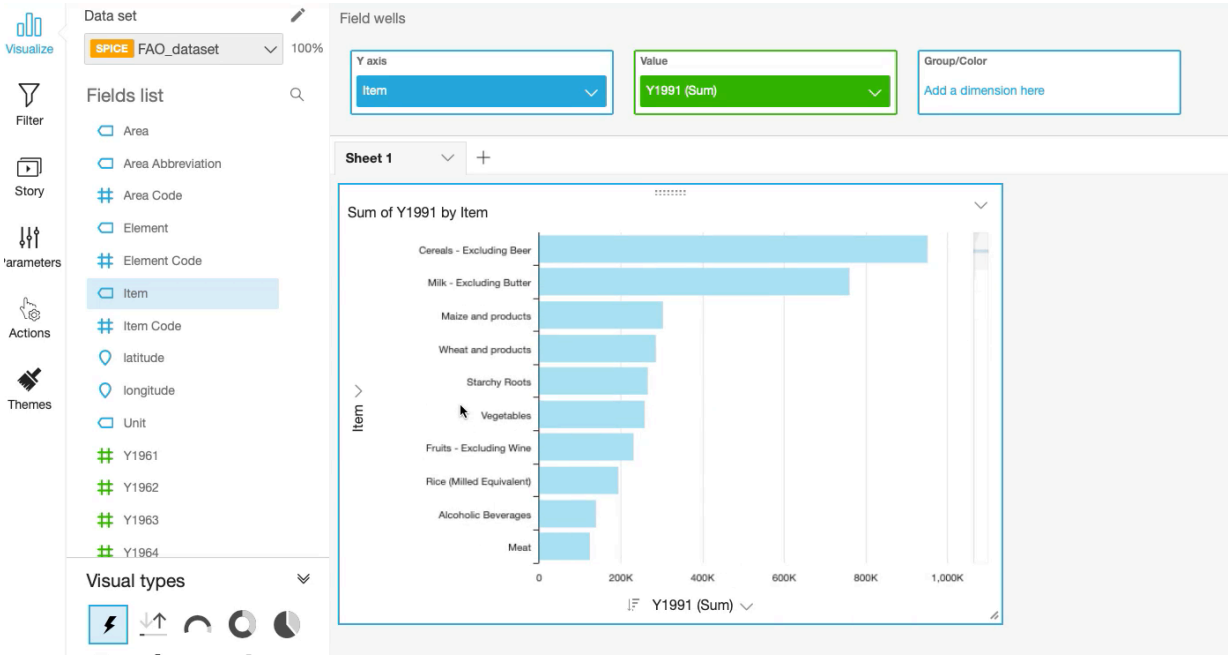

Finish data set creation
Table:
FAO_dataset
Estimated table size:
8.4MB
SPICE
Data source:
FAO_dataset
Import to SPICE
1023.4MB available
SPICE
Edit/Preview data
Visualize
S3



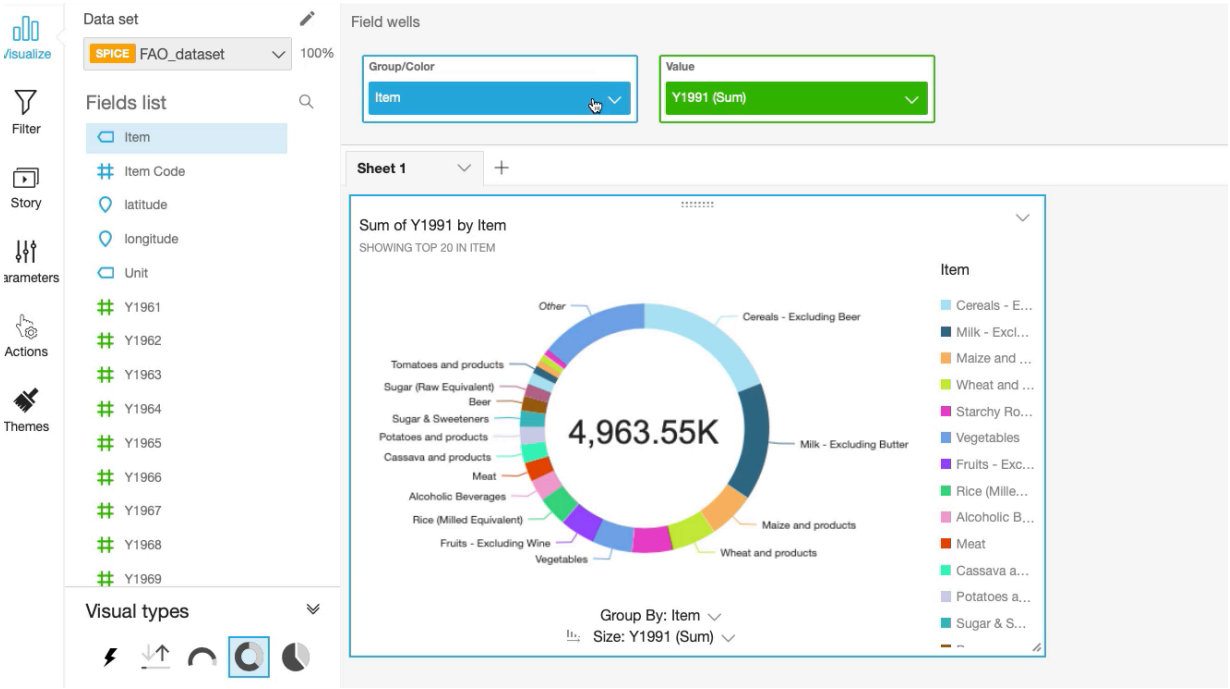
KPI example



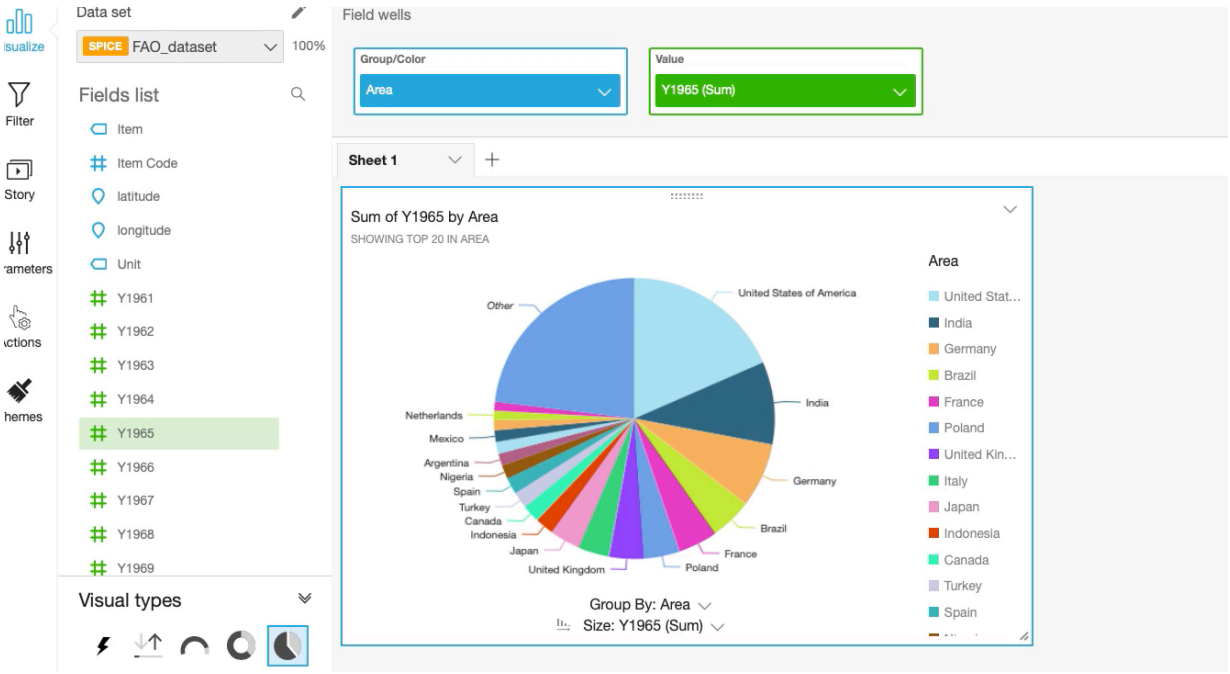
Auto-Graph



Doughnut example

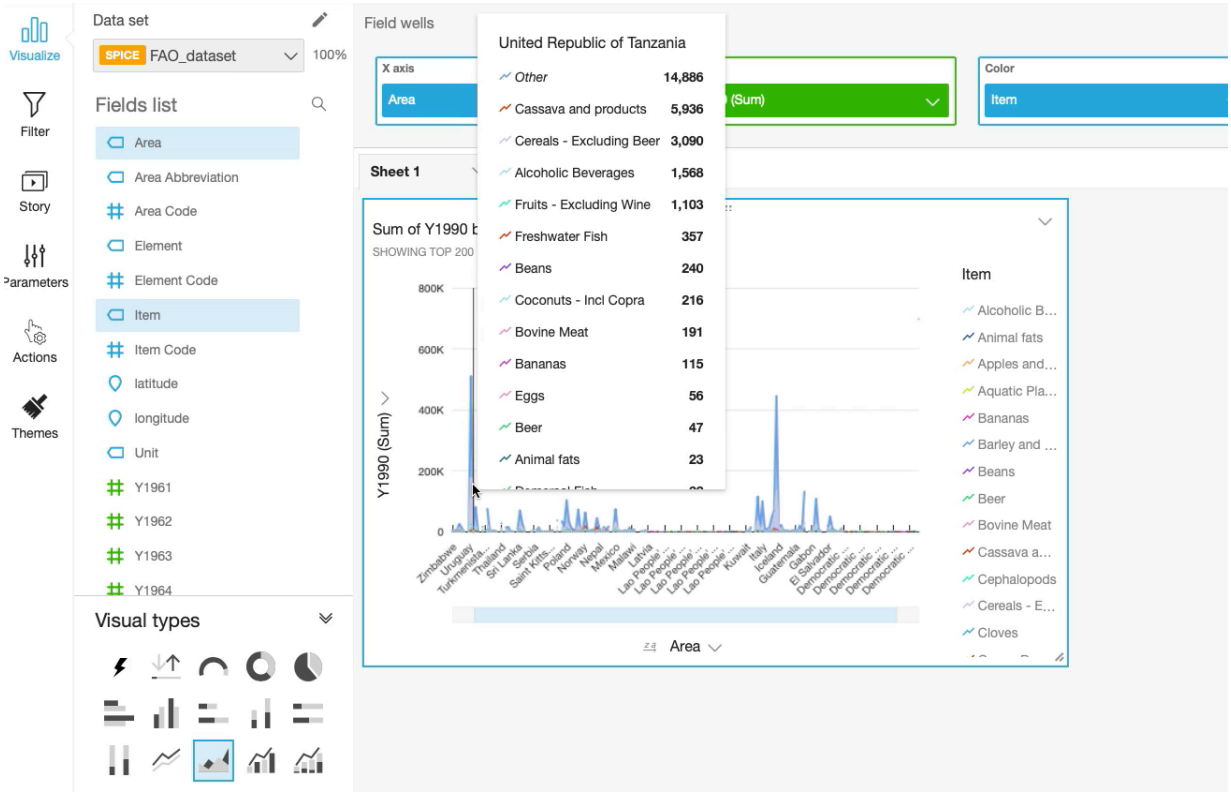


Pie chart example



Vertical 100% bar chart example

Area Line chart example



Pivot Table

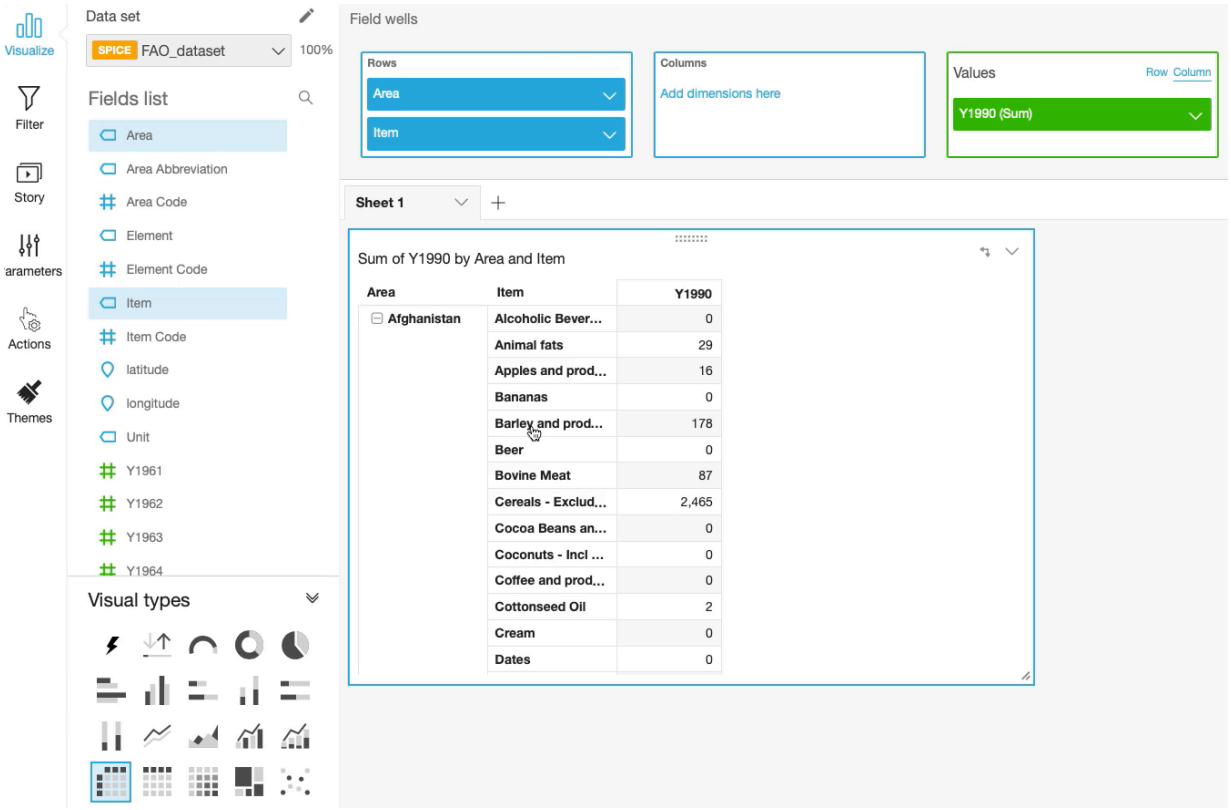
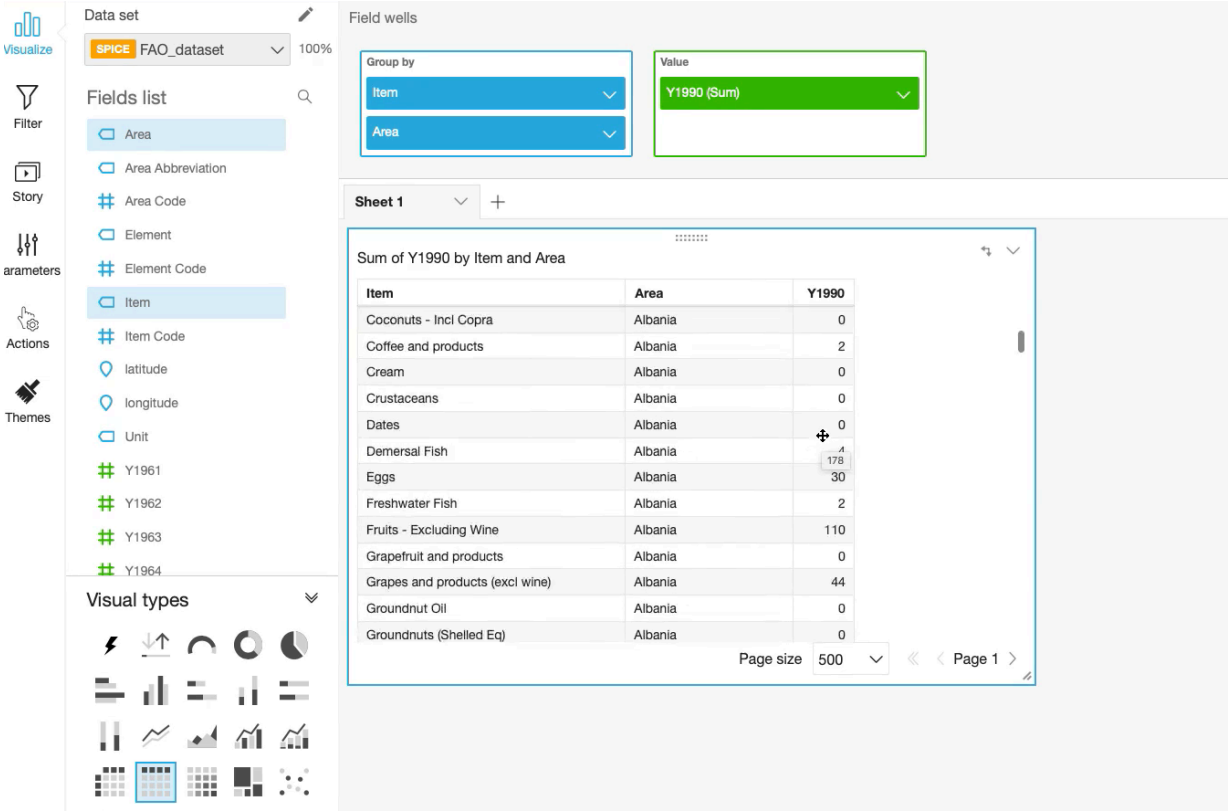
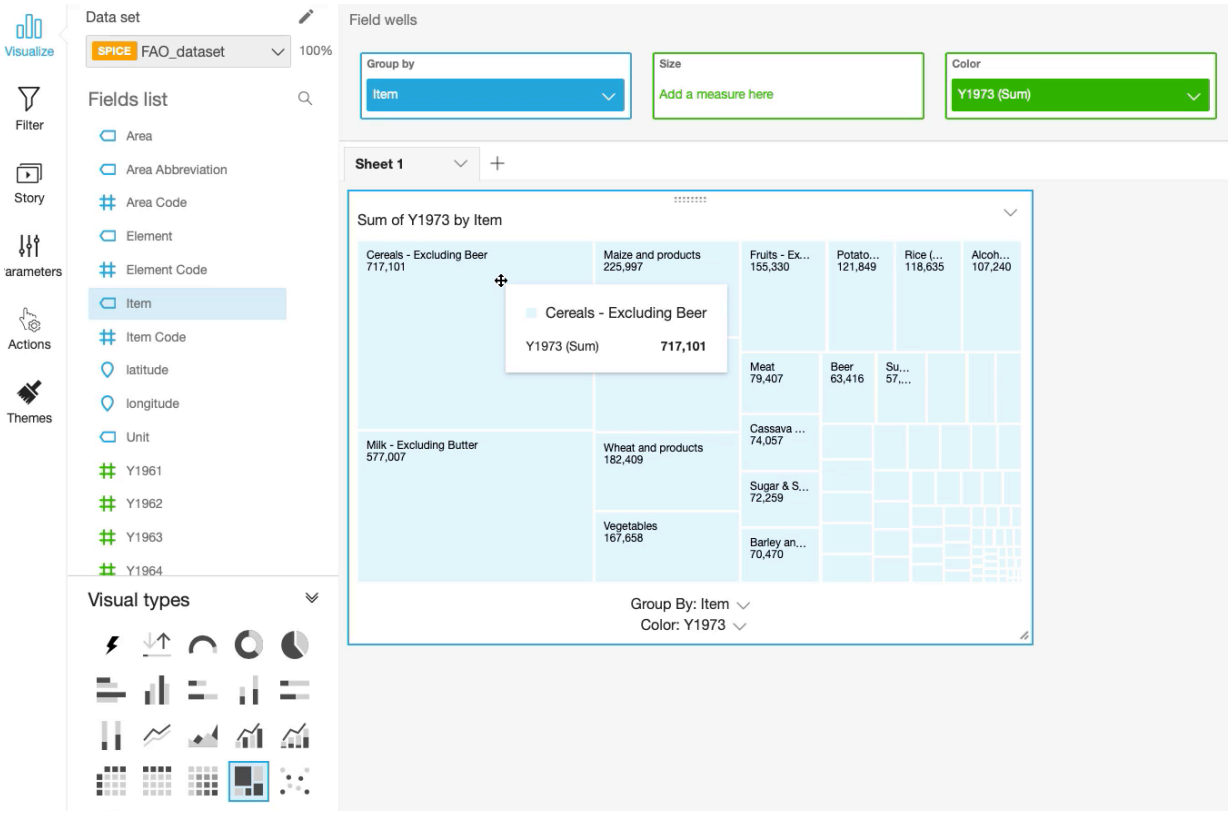


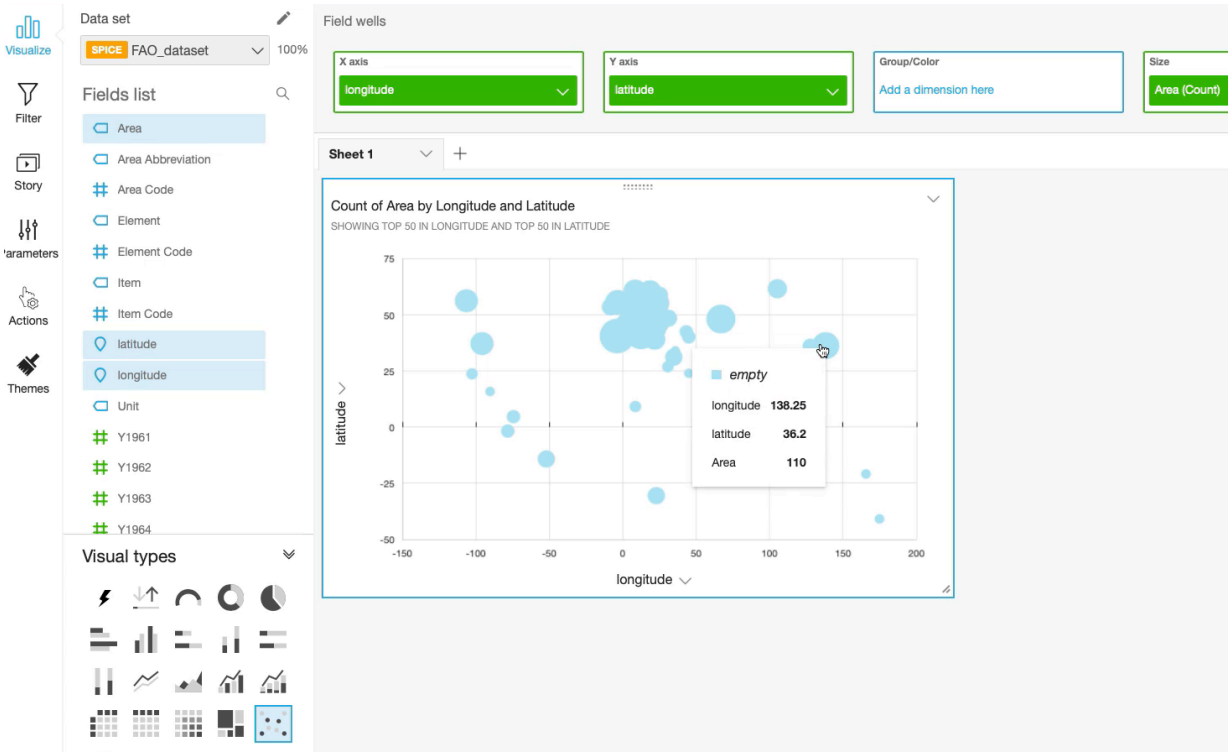
Table example



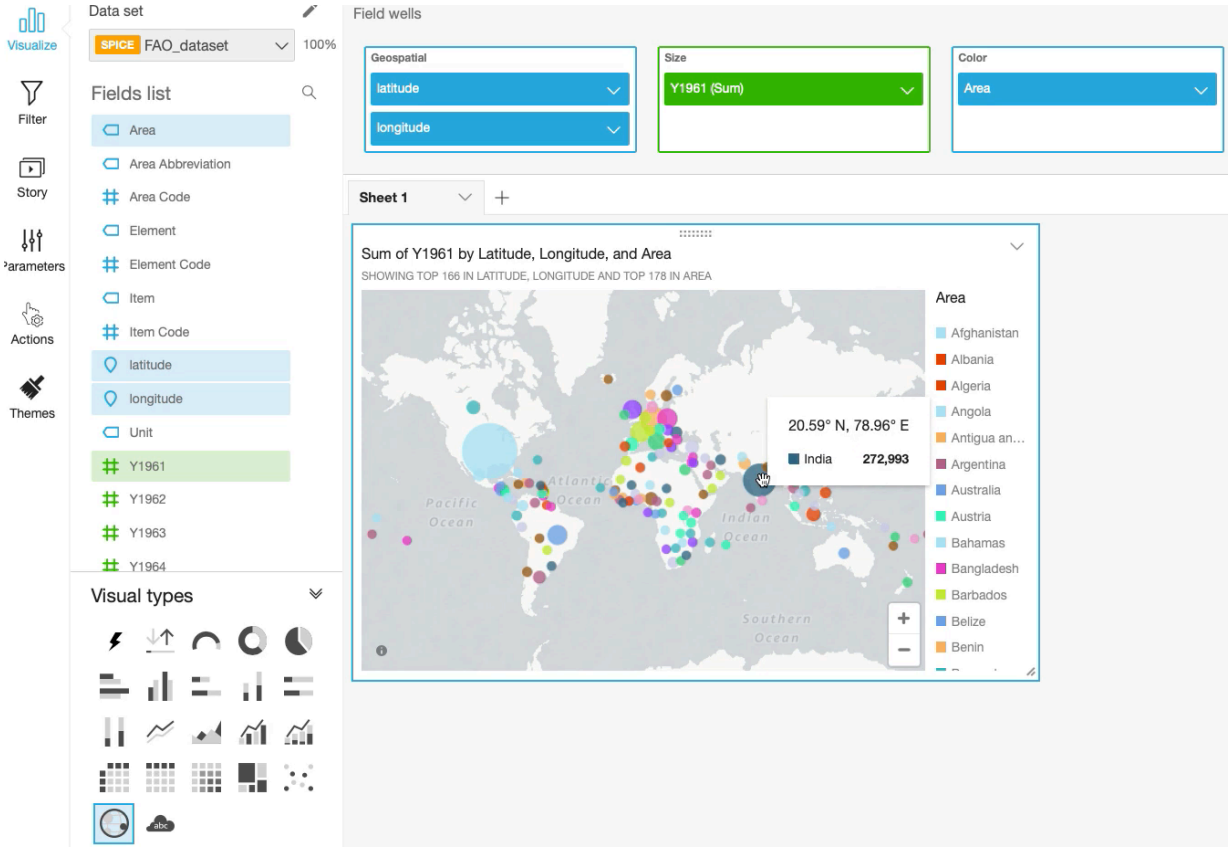
Tree Map example



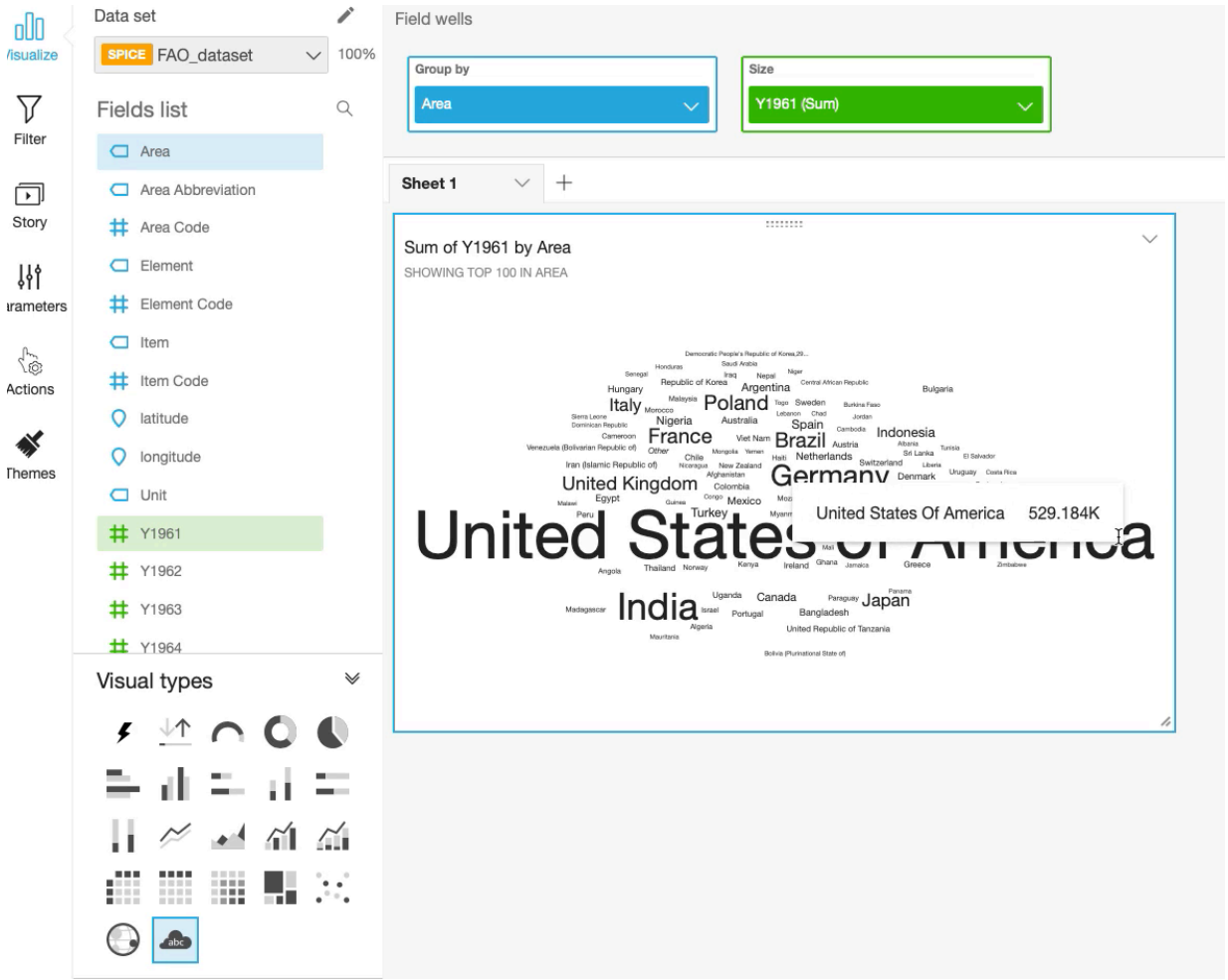
Scatter chart



Points in a map example



WordCloud example



Exploratory Data Analysis - Exam Tips

AWS Machine Learning Exploratory Data Analysis

Exam Tips - Data Streaming with Kinesis Data Streams

- ❏ Kinesis Data Streams
  - ❏ Continuously capture gigabytes of data per second from thousands of sources
  - ❏ Enables real-time analytics
  - ❏ Key Concepts
    - ❏ Data Producer
    - ❏ Data Consumer
    - ❏ Shard
    - ❏ Data Stream

Exam Tips - Data Streaming with Kinesis Data Firehose

- ❏ Kinesis Data Firehose
  - ❏ Fully managed service that automatically scales to match data throughput
  - ❏ Capture, transform, and load streaming data into S3, Redshift, Elasticsearch, and Splunk
  - ❏ Batch, compress, transform, and encrypt data before loading it onto your destination
  - ❏ Automatically convert incoming data to Apache Parquet/ORC before delivering to S3
  - ❏ Near real-time analytics with existing business intelligence tools
  - ❏ Requires no ongoing administration
- ❏ Key Concepts
  - ❏ Kinesis Data Delivery Stream
  - ❏ Elastic scaling handles variations in data throughput
  - ❏ Transform your data using Lambda

Exam Tips - Data Streaming with Kinesis Video Streams

- Kinesis Video Streams
  - Automatically provisions and scales infrastructure to read streaming media
  - Producers such as web cams, security cameras, audio feeds, images
    - Storage: Kinesis Video Streams ingests the stream data, stores, encrypts, and indexes the stream for either real-time or batch analytics
    - Consumers: Real-time or batch machine learning applications, Video processing or playback services
- Key Concepts
  - Used extensively in machine learning models

Exam Tips - Data Streaming with Kinesis Data Analytics

- Kinesis Data Analytics
  - Use SQL to process streaming data
  - Sources: Kinesis Data Streams and Kinesis Data Firehose
  - SQL queries put to S3, Redshift, or visualization and Business Intelligence tools
  - The streaming application continuously reads data from a streaming source, generates analytics using SQL code, and emits results to 1 to 4 destinations
- Key Concepts: the Streaming Application
  - Streaming application is the primary resource in Kinesis Data Analytics
  - A Kinesis Data Analytics streaming application continually reads and processes streaming data in real-time

Exam Tips - Data Transformation via AWS Glue

- ❏ AWS Glue
  - ❏ A fully managed ETL service for categorizing, cleaning, enriching, and moving your data
  - ❏ Glue components: Glue Catalog, ETL Engine, Scheduler
  - ❏ Serverless
  - ❏ Can convert semi-structured schemas to relational-schemas on the fly
  - ❏ Glue Terminology
- ❏ Key Concepts: the Glue components
  - ❏ Console: define and orchestrate ETL workflows
  - ❏ Data Catalog: persistent metadata store
  - ❏ Crawlers and Classifiers: crawlers scan data and classify it
  - ❏ ETL Operations: using metadata in the data catalog, autogenerates python or scala code
  - ❏ Jobs System: managed infrastructure to orchestrate your ETL workflow

Exam Tips - Analyzing and Visualizing Data for Machine Learning

- Types of information to convey via Business Intelligence (BI) tools
  - Key Performance Indicators (KPIs)
  - Relationships
  - Comparisons
  - Distributions
  - Compositions