

AWS ML - Data Analysis

1. Data Distributions
2. feature Engineering
 - Dealing with missing values
 - Dealing with unbalanced data
 - Handling Outliers
 - Numerical feature engineering
 - Transforming , Feature Extraction
 - Encoding Categorical features
 - Feature Selection
 - Shuffling
3. Amazon QuickSight
 - Features of Amazon Quicksights ML insights :
 - Quicksight Q
 - QuickSight Visual Types :
4. Elastic MapReduce (EMR)
5. Amazon Mechanical Turk

1. Data Distributions [↗](#)

- **Binomial** distributions are used for binary classifications of discrete events, such as flipping a coin.
- **Poisson** distributions for discrete events
- **Quidistributions** are generally for numerical data, not categorical.

2. feature Engineering [↗](#)

Dealing with missing values [↗](#)

- Do nothing
- **Most frequent value, Mean and Median replacement** : Median may be a better choice than mean when outliers are present. It's not the best solution : Only works on column level, misses correlations between features and Can't use on categorical features
- **Dropping**
- **MICE** Multiple Imputation by Chained Equations finds relationships between features and is one of the most advanced imputation methods available.
- **Using machine learning techniques** such as KNN, Regression
- **deep learning model** to impute missing values for categorical data

#todo : **Interpolation/Extrapolation, Forward Filling/Backword Filling, Hot deck Imputation**

Dealing with unbalanced data [↗](#)

Handling Outliers [↗](#)

Data points that lie more than one standard deviation (square root of variance) from the mean can be considered as unusual

AWS's Random Cut Forest algorithm for outlier detection. ,Found within QuickSight, Kinesis Analytics, SageMaker, and more

Numerical feature engineering [↗](#)

- **Normalization** : rescales values into [0,1]
- **Standardization** : rescales data to have a mean of 0 and a standard deviation of 1
- **Binning**
 - Transform numerical data to categorical data/ ordinal data when there is uncertainty in the measurements
 - Quantile binning : ensure having the same number of samples in each bin

Transforming , Feature Extraction [↗](#)

create or replace data by doing a transformation on it

Encoding Categorical features [↗](#)

- **One-hot encoding** for categorical data : create “buckets” for each category. 1 for a category and 0 for others
- **Binarizer Encoding** : features with binary nature
- **Label Encoding**
 - Ordinal Encoding

Feature Selection [↗](#)

Use feature selection to **filter irrelevant or redundant features**

- requires Normalization
- Removes features : **variance thresholds** (#todo)

Shuffling [↗](#)

Why? some models may learn from residual signals in the training data resulting from the order.

3. Amazon QuickSight [↗](#)

Business analytics and visualisations

Applications:

- Interactive viz
- Dashboards and KPI

Examples of data sources :

- Redshift
- Aurora/RDS
- EC2-hosted databases
- S3
- Snowflake

Data sets are imported into SPICE : Super-fast, Parallel, In-memory, Calculation Engine

Features of Amazon Quicksights ML insights : [↗](#)

- Anomaly detection
- forecasting

- auto-narratives

QuickSight Q

A feature in QuickSight that answers business questions **with natural language** instead of creating sql queries.

You must set up topics associated with datasets

QuickSight Visual Types :

- **AutoGraph** allows you to select the most appropriate visualisations
- **Bar Charts/Histograms** for comparison and distribution on a single dimension
- **Line graphs / Area line charts** : for changes over time.
- **Scatter plots/ heatmaps** for correlation, multiple distribution
- **Pie graphs** for aggregation, categorical attributes
- **Donut charts** : Pourcentage of Total amount
- **Gauge Charts** : Compare values in a measure
- **Tree maps** for hierarchical aggregation,
- **Pivot tables** for tabular data
- **KPI's** : compare key value to its target value
- **Geospatial charts**
- **World Clouds** : word or phrase frequency
- **Pair plots** are best used for spotting correlations between pairs of attributes.
- **Box & whisker**, or just "box plots", organize your data into quartiles, and display outliers in the outer quartiles, distribution on a multiple dimensions

4. Elastic MapReduce (EMR)

- Managed Hadoop framework on EC2 instances
- Includes Spark, HBase, Presto, Flink, Hive and more
- If you have a massive dataset that you need to preprocess, EMR provides a way to distributing the load of processing that data across an entire cluster of computers
- A cluster in EMR is a collection of nodes (EC2 instances)
 - **Master node:** manages the cluster
 - Single EC2 instance
 - **Core node:** Hosts HDFS data and runs tasks
 - Can be scaled up & down, but with some risk
 - **Task node:** Runs tasks, does not host data
 - No risk of data loss when removing
 - Good use of **spot instances**

EMR / AWS Integration

- Amazon EC2 for the instances that comprise the nodes in the cluster
- Amazon VPC to configure the virtual network in which you launch your instances
- Amazon S3 to store input and output data
- Amazon CloudWatch to monitor cluster performance and configure alarms
- AWS IAM to configure permissions
- AWS CloudTrail to audit requests made to the service

- AWS Data Pipeline to schedule and start your clusters

EMR Storage options

- HDFS
- EMRFS: access S3 as if it were HDFS
- Local file system
- EBS (Elastic Block Storage) for HDFS

5. Amazon Mechanical Turk [↗](#)