

Tech experts

3ª semana

Data Visualization

- Visualizing relationships in your data
 - Scatter plot
 - Bubble chart

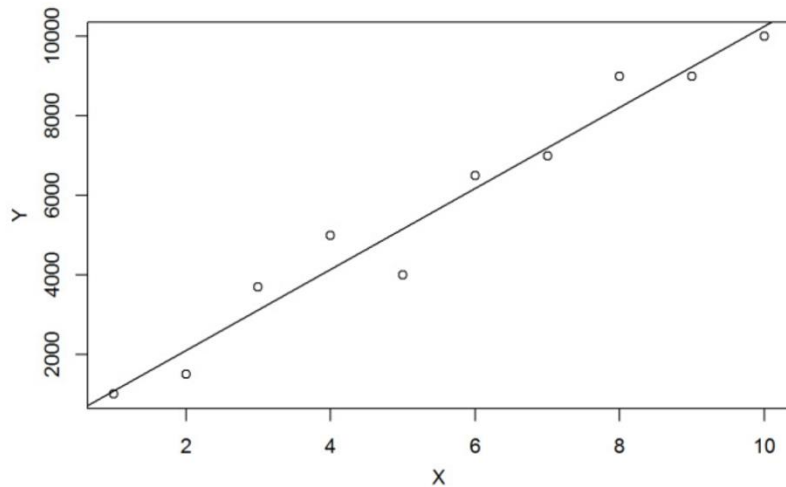


Figure 4.1 – Plotting relationships with a scatter plot

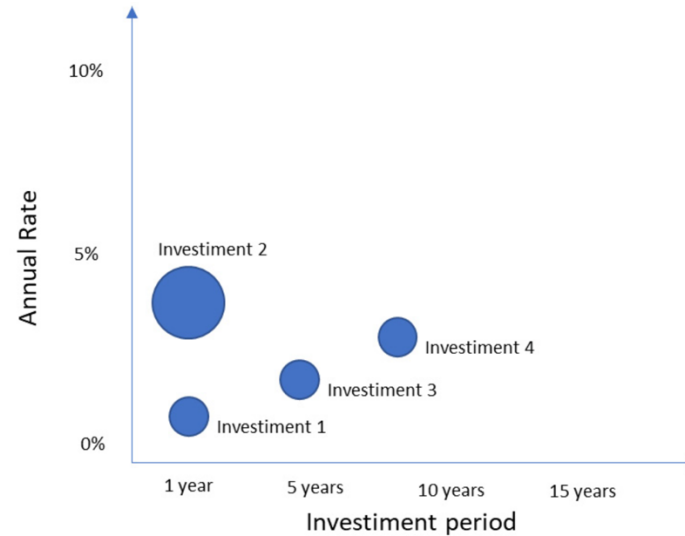


Figure 4.2 – Plotting relationships with a bubble chart

Data Visualization

- Visualizing comparisons in your data
 - Bar chart
 - stacked column charts
 - Column chart
 - Line chart

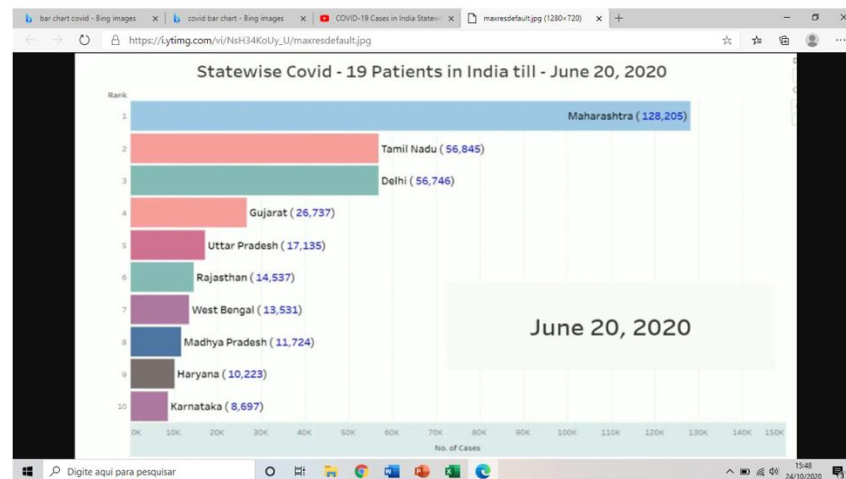


Figure 4.3 – Plotting comparisons with a bar chart

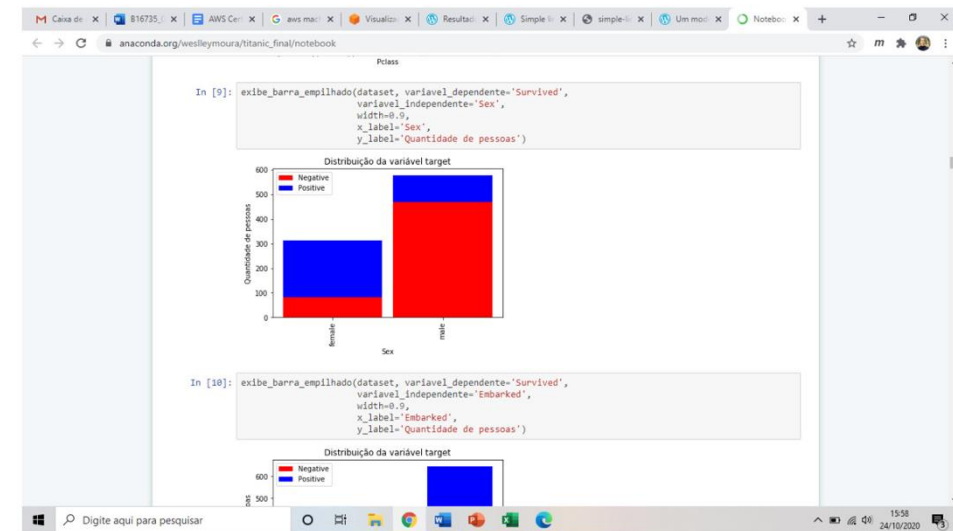


Figure 4.4 – Using a stacked bar chart to analyze the Titanic disaster dataset

Data Visualization

- Visualizing comparisons in your data
 - Bar chart
 - stacked column charts
 - Column chart
 - Line chart

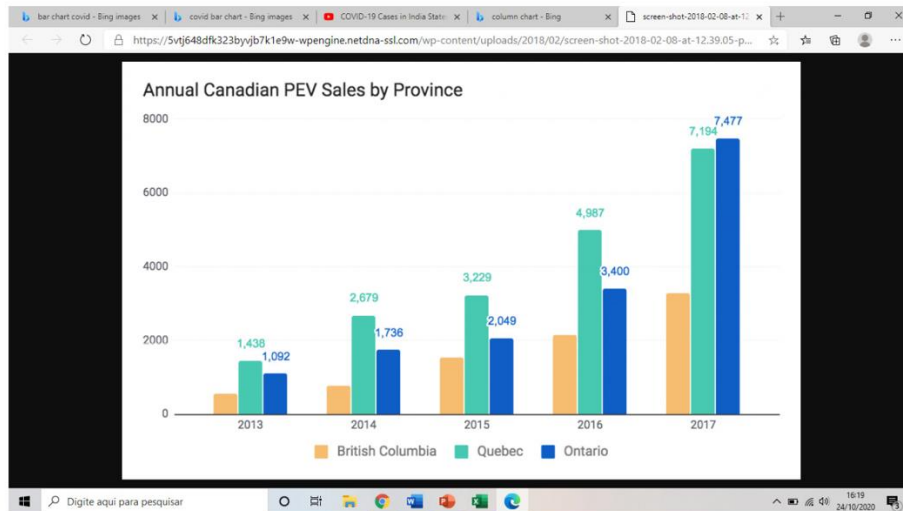


Figure 4.5 – Plotting comparisons with a column chart

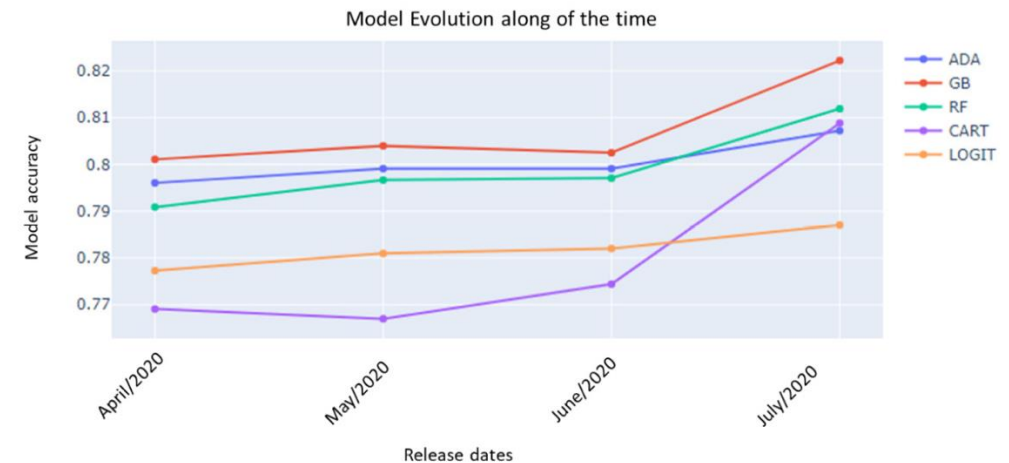


Figure 4.6 – Plotting comparisons with a line chart

Data Visualization

- Visualizing distributions in your data
 - Histogram
 - Box plot

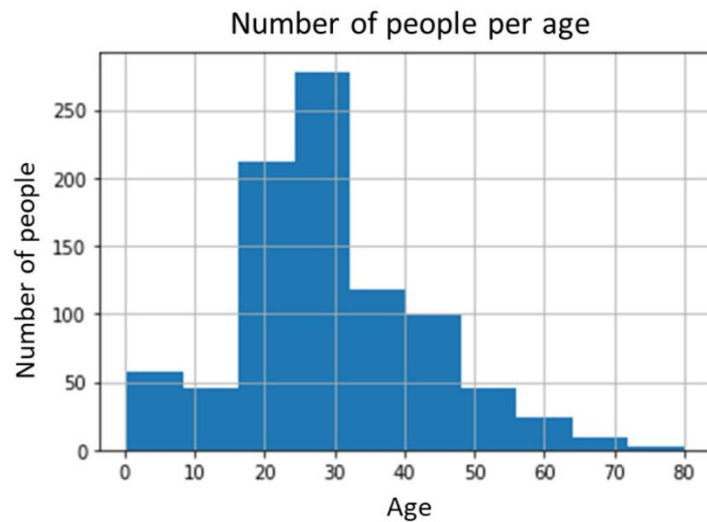


Figure 4.7 – Plotting distributions with a histogram

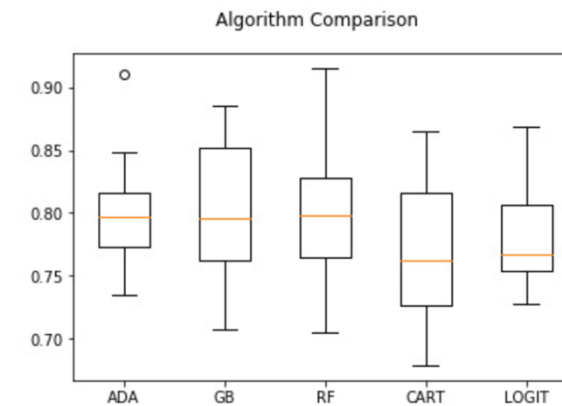


Figure 4.9 – Plotting distributions with a box plot

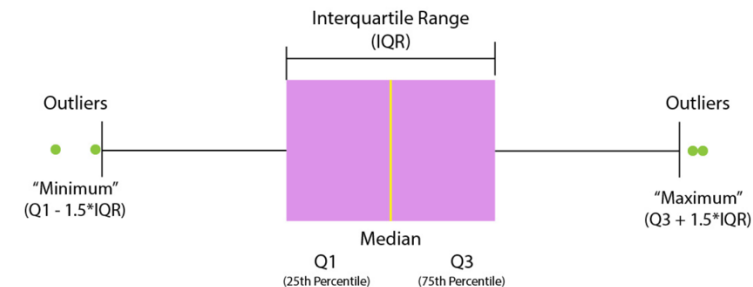
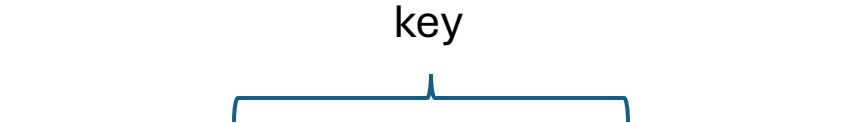


Figure 4.10 – Box plot elements

AWS Services for Data Storing

- S3
 - Buckets
 - Objects
 - Key
 - Value(max size: 5TB)
 - S3 object URI: s3://bucket-name/folder-as-a-prefix/image.jpg
 - Storage classes:
 - <https://aws.amazon.com/pt/s3/storage-classes>
 - Bucket policy
 - Bucker versioning

AWS Services for Data Storing

- S3
 - Encryption
 - **Client-Side Encryption: 100% gerenciado pelo cliente**
 - **Server-Side Encryption with Customer-Provided Keys (SSE-C):** Cliente fornece a chave no momento de gravar e ler o objeto, mas AWS faz a operação de criptografia
 - **Server-Side Encryption with Amazon S3-Managed Keys (SSE-S3):** AWS cuida das chaves e da operação. **Padrão.**
 - **Server-Side Encryption with Customer Master Keys stored in AWS Key Management Service (SSE-KMS):** KMS gerencia a chave, S3 realizar a operação. Cliente pode definir regras de permissionamento específicos para as chaves, além de rotação das chaves.

AWS Services for Data Storing

- **Elastic Block Store (EBS)**
- Volumes atachados à EC2.
 - Pode ser criptografado com chaves KMS.
 - != Instance Store.

Volume Types	Use cases
General Purpose SSD (gp2)	Useful for maintaining balance between price and performance. Good for most workloads, system boot volumes, dev, and test environments.
Provisioned IOPS SSD (io2, io1)	Useful for mission-critical, high-throughput or low-latency workloads. For example, I/O intensive database workloads like MongoDB, Cassandra, Oracle
Throughput Optimized HDD (st1)	Useful for frequently accessed, throughput-intensive workloads. For example, big data processing, data warehouses, log processing
Cold HDD (sc1)	Useful for less frequently accessed workloads.

Figure 5.3 – Different volumes and their use cases

AWS Services for Data Storing

- **Elastic File System (EFS)**

- Network based filesystem
- Linux EC2
- Pode ser utilizado por várias máquinas ao mesmo tempo
- Se máquina desligar, não se perde os dados.
- On-premises podem acessar utilizando VPN ou Direct Connect.

Question 1

- A Data Engineer needs to build a model using a dataset containing customer credit card information. How can the Data Engineer ensure the data remains encrypted and the credit card information is secure?
- A. Use a custom encryption algorithm to encrypt the data and store the data on an Amazon SageMaker instance in a VPC. Use the SageMaker DeepAR algorithm to randomize the credit card numbers.
- B. Use an IAM policy to encrypt the data on the Amazon S3 bucket and Amazon Kinesis to automatically discard credit card numbers and insert fake credit card numbers.
- C. Use an Amazon SageMaker launch configuration to encrypt the data once it is copied to the SageMaker instance in a VPC. Use the SageMaker principal component analysis (PCA) algorithm to reduce the length of the credit card numbers.
- D. Use AWS KMS to encrypt the data on Amazon S3 and Amazon SageMaker, and redact the credit card numbers from the customer data with AWS Glue.

Question 1

- A Data Engineer needs to build a model using a dataset containing customer credit card information. How can the Data Engineer ensure the data remains encrypted and the credit card information is secure?
- A. Use a custom encryption algorithm to encrypt the data and store the data on an Amazon SageMaker instance in a VPC. Use the SageMaker DeepAR algorithm to randomize the credit card numbers.
- B. Use an IAM policy to encrypt the data on the Amazon S3 bucket and Amazon Kinesis to automatically discard credit card numbers and insert fake credit card numbers.
- C. Use an Amazon SageMaker launch configuration to encrypt the data once it is copied to the SageMaker instance in a VPC. Use the SageMaker principal component analysis (PCA) algorithm to reduce the length of the credit card numbers.
- D. Use AWS KMS to encrypt the data on Amazon S3 and Amazon SageMaker, and redact the credit card numbers from the customer data with AWS Glue.

Question 2

- A Machine Learning Specialist is using an Amazon SageMaker notebook instance in a private subnet of a corporate VPC. The ML Specialist has important data stored on the Amazon SageMaker notebook instance's Amazon EBS volume, and needs to take a snapshot of that EBS volume. However, the ML Specialist cannot find the Amazon SageMaker notebook instance's EBS volume or Amazon EC2 instance within the VPC.
Why is the ML Specialist not seeing the instance visible in the VPC?
- A. Amazon SageMaker notebook instances are based on the EC2 instances within the customer account, but they run outside of VPCs.
- B. Amazon SageMaker notebook instances are based on the Amazon ECS service within customer accounts.
- C. Amazon SageMaker notebook instances are based on EC2 instances running within AWS service accounts.
- D. Amazon SageMaker notebook instances are based on AWS ECS instances running within AWS service accounts.

Question 2

- A Machine Learning Specialist is using an Amazon SageMaker notebook instance in a private subnet of a corporate VPC. The ML Specialist has important data stored on the Amazon SageMaker notebook instance's Amazon EBS volume, and needs to take a snapshot of that EBS volume. However, the ML Specialist cannot find the Amazon SageMaker notebook instance's EBS volume or Amazon EC2 instance within the VPC. Why is the ML Specialist not seeing the instance visible in the VPC?
- A. Amazon SageMaker notebook instances are based on the **EC2 instances within the customer account, but they run outside of VPCs.**
- B. Amazon SageMaker notebook instances are based on the **Amazon ECS** service within customer accounts.
- **C. Amazon SageMaker notebook instances are based on EC2 instances running within AWS service accounts.**
- D. Amazon SageMaker notebook instances are based on **AWS ECS instances** running within AWS service accounts.

Question 3

- A Machine Learning Specialist at a company sensitive to security is preparing a dataset for model training. The dataset is stored in Amazon S3 and contains Personally Identifiable Information (PII). The dataset:
 - ☞ Must be accessible from a VPC only.
 - ☞ Must not traverse the public internet.How can these requirements be satisfied?
- A. Create a VPC endpoint and apply a bucket access policy that restricts access to the given VPC endpoint and the VPC.
- B. Create a VPC endpoint and apply a bucket access policy that allows access from the given VPC endpoint and an Amazon EC2 instance.
- C. Create a VPC endpoint and use Network Access Control Lists (NACLs) to allow traffic between only the given VPC endpoint and an Amazon EC2 instance.
- D. Create a VPC endpoint and use security groups to restrict access to the given VPC endpoint and an Amazon EC2 instance

Question 3

- A Machine Learning Specialist at a company sensitive to security is preparing a dataset for model training. The dataset is stored in Amazon S3 and contains Personally Identifiable Information (PII). The dataset:
 - ☞ Must be accessible from a VPC only.
 - ☞ Must not traverse the public internet.How can these requirements be satisfied?
- A. Create a VPC endpoint and apply a bucket access policy that restricts access to the given VPC endpoint and the VPC.
- B. Create a VPC endpoint and apply a bucket access policy that allows access from the given VPC endpoint and an Amazon EC2 instance.
- C. Create a VPC endpoint and use Network Access Control Lists (NACLs) to allow traffic between only the given VPC endpoint and an Amazon EC2 instance.
- D. Create a VPC endpoint and use security groups to restrict access to the given VPC endpoint and an Amazon EC2 instance.

Question 4

- A Machine Learning Specialist at a company sensitive to security is preparing a dataset for model training. The dataset is stored in Amazon S3 and contains Personally Identifiable Information (PII). The dataset:
 - ☞ Must be accessible from a VPC only.
 - ☞ Must not traverse the public internet.How can these requirements be satisfied?
- A. Create a VPC endpoint and apply a bucket access policy that restricts access to the given VPC endpoint and the VPC.
- B. Create a VPC endpoint and apply a bucket access policy that allows access from the given VPC endpoint and an Amazon EC2 instance.
- C. Create a VPC endpoint and use Network Access Control Lists (NACLs) to allow traffic between only the given VPC endpoint and an Amazon EC2 instance.
- D. Create a VPC endpoint and use security groups to restrict access to the given VPC endpoint and an Amazon EC2 instance

Question 4

- A Machine Learning Specialist at a company sensitive to security is preparing a dataset for model training. The dataset is stored in Amazon S3 and contains Personally Identifiable Information (PII). The dataset:
 - ☞ Must be accessible from a VPC only.
 - ☞ Must not traverse the public internet.How can these requirements be satisfied?
- A. Create a VPC endpoint and apply a bucket access policy that restricts access to the given VPC endpoint and the VPC.
- B. Create a VPC endpoint and apply a bucket access policy that allows access from the given VPC endpoint and an Amazon EC2 instance.
- C. Create a VPC endpoint and use Network Access Control Lists (NACLs) to allow traffic between only the given VPC endpoint and an Amazon EC2 instance.
- D. Create a VPC endpoint and use security groups to restrict access to the given VPC endpoint and an Amazon EC2 instance.

Question 5

- A Data Science team is designing a dataset repository where it will store a large amount of training data commonly used in its machine learning models. As Data Scientists may create an arbitrary number of new datasets every day, the solution has to scale automatically and be cost-effective. Also, it must be possible to explore the data using SQL.
Which storage scheme is MOST adapted to this scenario?
- A. Store datasets as files in Amazon S3.
- B. Store datasets as files in an Amazon EBS volume attached to an Amazon EC2 instance.
- C. Store datasets as tables in a multi-node Amazon Redshift cluster.
- D. Store datasets as global tables in Amazon DynamoDB.

Question 5

- A Data Science team is designing a dataset repository where it will store a large amount of training data commonly used in its machine learning models. As Data Scientists may create an arbitrary number of new datasets every day, the solution has to scale automatically and be **cost-effective**. Also, it must be possible to explore the data using SQL.
Which storage scheme is MOST adapted to this scenario?
- A. **Store datasets as files in Amazon S3.**
- B. Store datasets as files in an **Amazon EBS volume attached to an Amazon EC2 instance.**
- C. Store datasets as tables in a multi-node Amazon Redshift cluster.
- D. Store datasets **as global tables in Amazon DynamoDB.**