Installing Packages & Environment Setup

```
!pip install pandas-gbq --quiet
!pip install google-cloud-bigquery pandas
!pip install --quiet google-cloud-bigquery
from google.colab import auth
auth.authenticate_user()
import pandas as pd
from pandas.io import gbq
import matplotlib.pyplot as plt
import seaborn as sns
import plotly.express as px
from google.cloud import bigquery
```

Show hidden output

Executive Summary

This project explores the Bitcoin Cash blockchain by analyzing block-level data using BigQuery. Our primary objective was to build a predictive model that determines whether a block exceeds 500KB in size based on metadata such as block version, number, and time since the previous block.

After cleaning and preparing 5,000 rows of blockchain data, we trained a Decision Tree Classifier. The model achieved strong performance with an overall test accuracy of 91%.

Key results:

- Precision (for large blocks): 0.79
- Recall (for large blocks): 0.80
- F1-score (for large blocks): 0.79

The model demonstrates solid predictive power and highlights potential for deeper analysis on blockchain behavior using machine learning techniques. Future improvements could include tuning model hyperparameters, testing ensemble models, or analyzing trends across time windows.

Project Connection

```
project_id = 'proven-wavelet-457219-u4'
client = bigquery.Client(project = project_id)
```

Dataset Description

The dataset used in this project is sourced from BigQuery's public dataset: `bigquery-public-data.crypto_bitcoin_cash.blocks`.

Each row in the dataset represents a block on the Bitcoin Cash blockchain. The following variables were selected for modeling:

- size: The size of the block in bytes (used to create the target variable).
- version: The block version number.
- number: The block height (its position in the chain).
- nonce: A value miners vary to find a valid hash (excluded from final model).
- time_since_last_block: Time difference (in seconds) between this and the previous block.

The target variable `label` was engineered as a binary indicator:

- 1 if the block size > 500,000 bytes
- 0 otherwise

After filtering and cleaning, the final dataset contains 5,000 rows and 4 predictor features. This dataset was suitable for a binary classification task.

```
query = """
SELECT *
```

```
FROM `bigquery-public-data.crypto_bitcoin_cash.blocks`
LIMIT 10
"""

result = client.query(query).result().to_dataframe()
result.head()
```

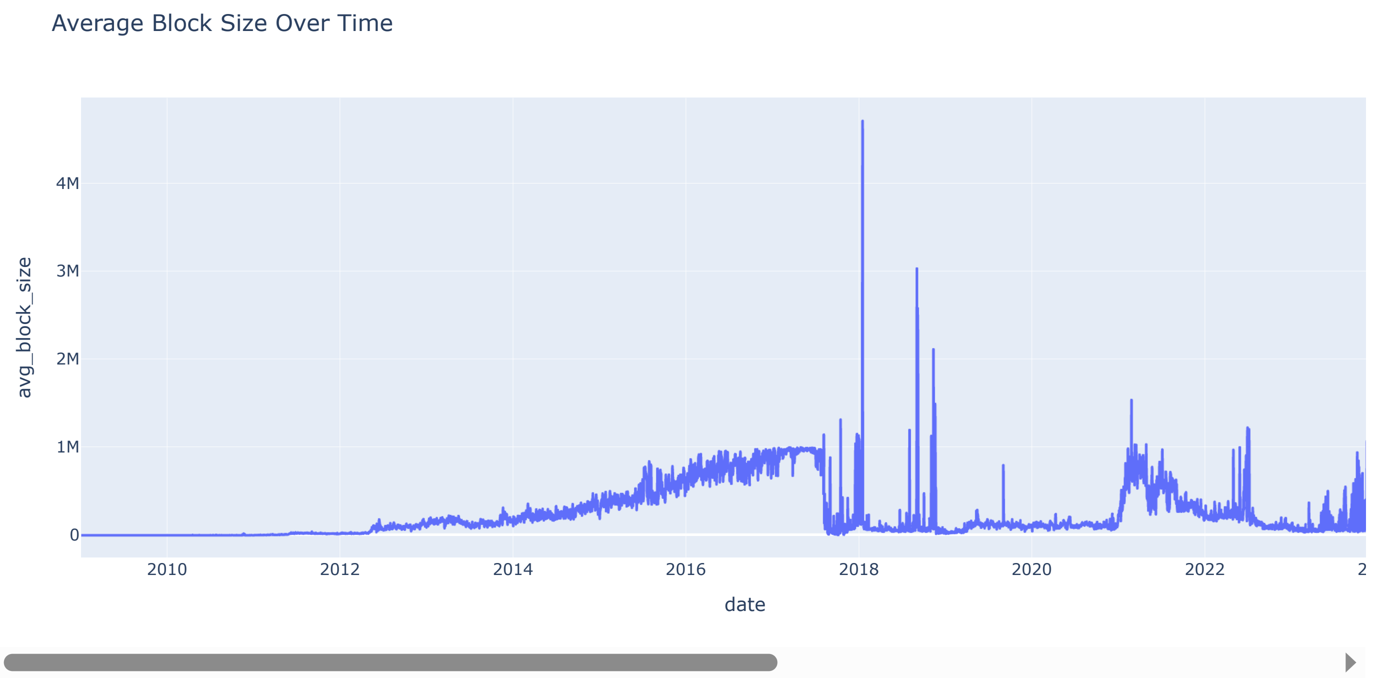
Next steps: Generate code with result View recommended plots New interactive sheet

EDA RESULTS and VISUALS

Query 1: Block sizes over time

```
query1 = """
SELECT
  DATE(timestamp) AS date,
  AVG(size) AS avg_block_size
FROM `bigquery-public-data.crypto_bitcoin_cash.blocks`
GROUP BY date
ORDER BY date
"""
df1 = client.query(query1).result().to_dataframe()


px.line(df1, x='date', y='avg_block_size', title='Average Block Size Over Time').show()
```



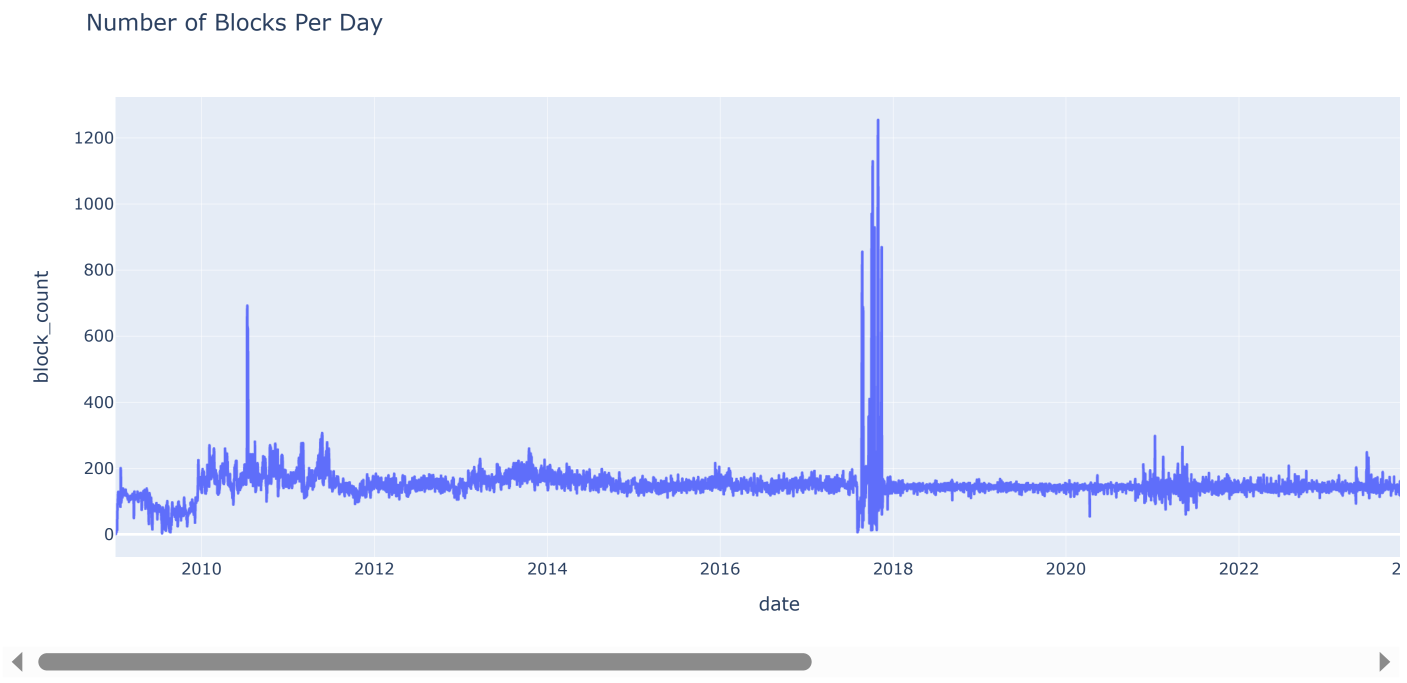
The average Bitcoin block size has generally increased over time, especially from around 2013 to 2017, indicating greater transaction volume and network activity. However, there are noticeable sharp drops and volatility after 2017, suggesting changes in network behavior, possibly due to scaling solutions (like SegWit) or market events impacting transaction patterns

Query 2: Number of blocks per day

```
query2 = """
SELECT
  DATE(timestamp) AS date,
  COUNT(*) AS block_count
FROM `bigquery-public-data.crypto_bitcoin_cash.blocks`
```

```
GROUP BY date
ORDER BY date
"""
df2 = client.query(query2).result().to_dataframe()


px.line(df2, x='date', y='block_count', title='Number of Blocks Per Day').show()
```



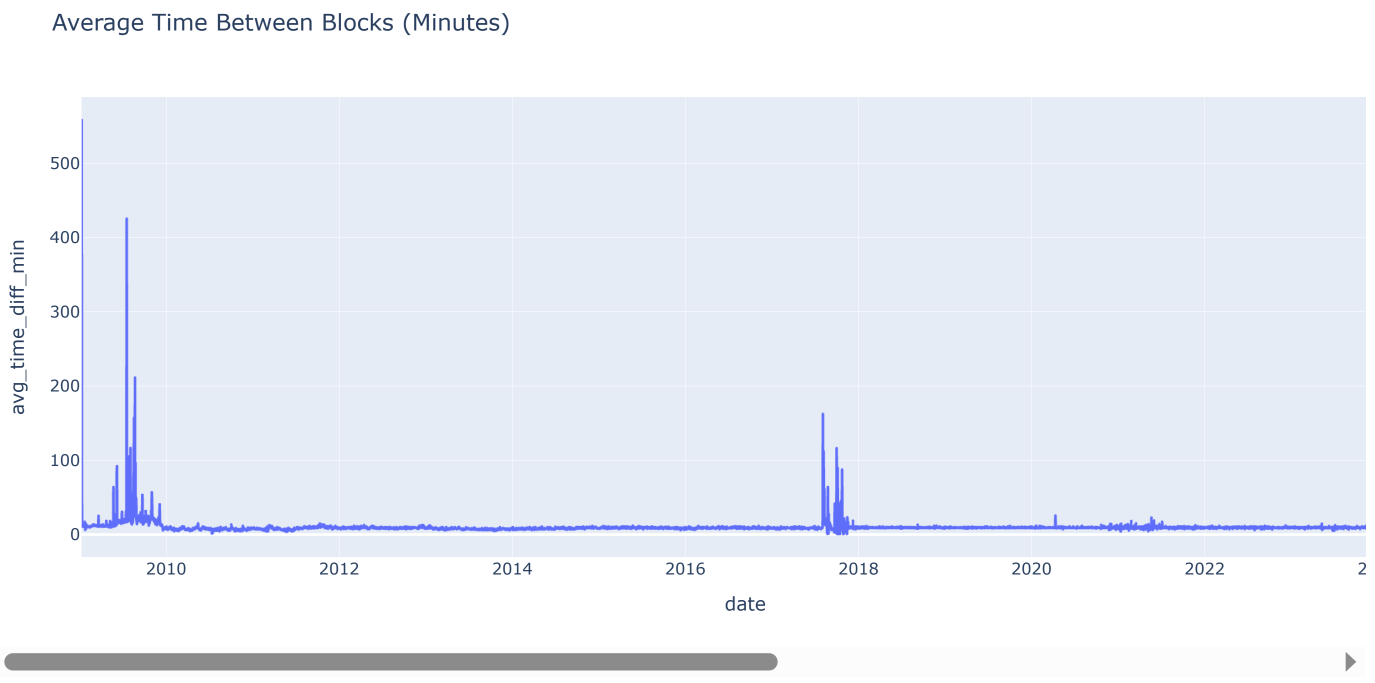
The number of Bitcoin blocks mined per day has remained relatively stable over time, hovering around 150-200 blocks/day. However, there are occasional spikes, most notably around late 2017 to early 2018, when the number sharply increased to over 1200 blocks/day. This spike may reflect abnormal network behavior or testing phases, possibly linked to a hard fork, stress test, or sudden difficulty adjustments.

Query 3: Average time between blocks

```
query3 = """
WITH block_times AS (
  SELECT
    timestamp,
    TIMESTAMP_DIFF(timestamp, LAG(timestamp) OVER (ORDER BY timestamp), MINUTE) AS time_diff_min
  FROM
    `bigquery-public-data.crypto_bitcoin_cash.blocks`
)

SELECT
  DATE(timestamp) AS date,
  AVG(time_diff_min) AS avg_time_diff_min
FROM block_times
WHERE time_diff_min IS NOT NULL
GROUP BY date
ORDER BY date
"""
df3 = client.query(query3).result().to_dataframe()


px.line(df3, x='date', y='avg_time_diff_min', title='Average Time Between Blocks (Minutes)').show()
```



The average time between Bitsoin blocks has remained fairly consistent around the expected target of 10 min, especially after 2010. However, there were a few significant spikes--most notably before 2010 and around 2017 -- where the time between blocks dramatically increased. These spikes may indicate early network instability or difficulty adjustment issues during periods of extreme mining activity or reduced hash rate.

Predictive Modeling

Query and Prepare the Data

```
query = """
WITH block_data AS (
  SELECT
    size,
    weight,
    version,
    number,
    nonce,
    TIMESTAMP_DIFF(timestamp, LAG(timestamp) OVER (ORDER BY timestamp), SECOND) AS time_since_last_block
  FROM `bigquery-public-data.crypto_bitcoin_cash.blocks`
)
SELECT *
FROM block_data
WHERE time_since_last_block IS NOT NULL
  AND size IS NOT NULL

LIMIT 5000
"""
df = client.query(query).result().to_dataframe()
df = df.drop(columns=['weight'])
df = df.dropna()
df['label'] = (df['size'] > 500000).astype(int)
print(df.shape)
print(df['label'].value_counts())
```

Train/Test Split and Feature Prep

```
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler

# Features and target
X = df[[ 'version', 'number',  'time_since_last_block']]
```

```
y = df['label']

# Split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# Scale features
scaler = StandardScaler()
X_train_scaled = scaler.fit_transform(X_train)
X_test_scaled = scaler.transform(X_test)
```

Train & Evaluate a Decision Tree Classifier

```
from sklearn.tree import DecisionTreeClassifier
from sklearn.metrics import classification_report, confusion_matrix, ConfusionMatrixDisplay

# Initialize and train the model
tree_model = DecisionTreeClassifier(random_state=42)
tree_model.fit(X_train_scaled, y_train)

# Predict
y_pred = tree_model.predict(X_test_scaled)

# Evaluate
print(confusion_matrix(y_test, y_pred))
print(classification_report(y_test, y_pred))

# Optional: Visualize confusion matrix
ConfusionMatrixDisplay.from_predictions(y_test, y_pred)
```

Model Evaluation Summary

We trained a Decision Tree Classifier to predict whether a block size exceeds 500KB. The model achieved an accuracy of approximately 91% on the test set.")

Key metrics:

- Precision (Class 1): 0.79 — Of all blocks predicted as large, 79% were actually large.
- Recall (Class 1): 0.80 — The model correctly identified 80% of actual large blocks.
- F1-score (Class 1): 0.79 — Balanced performance on precision and recall for large blocks.

The confusion matrix shows:

- True Negatives: 733
- False Positives: 47
- False Negatives: 45
- True Positives: 175

The model performs well overall, with stronger accuracy on the majority class.") It can be improved by trying techniques like hyperparameter tuning or using ensemble models.

Managerial Insights and Takeaways

1. **Blockchain block size is predictable using metadata:** Variables like block version, position (number), and time between blocks offer meaningful signals that can help anticipate whether a block will be large.
2. **Machine learning can effectively support blockchain analysis:** The decision tree model achieved 91% accuracy, suggesting that predictive models can be used for monitoring, optimization, or anomaly detection in blockchain operations.
3. **Operational planning opportunities for network scalability:** Knowing which blocks are likely to be large may help miners or network operators better manage bandwidth, node performance, and transaction prioritization.
4. **Model interpretability supports decision-making:** The decision tree model is transparent, making it easier to explain to stakeholders and adapt into rule-based systems or dashboards.
5. **Data from public sources like BigQuery can power real insights:** This project demonstrates how publicly available blockchain data can be leveraged for business intelligence and innovation.