Using BigQuery ML to Predict Penguin Weight

2 hoursFree

Overview

In this lab, you use the `penguin table` to create a model that predicts the weight of a penguin based on the penguin's species, island of residence, culmen length and depth, flipper length, and sex.
This lab introduces data analysts to BigQuery ML. BigQuery ML enables users to create and execute machine learning models in BigQuery using SQL queries. The goal is to democratize machine learning by enabling SQL practitioners to build models using their existing tools and to increase development speed by eliminating the need for data movement.

Learning objectives

- Create a linear regression model using the `CREATE MODEL` statement with BigQuery ML.

- Evaluate the ML model with the `ML.EVALUATE` function.
- Make predictions using the ML model with the `ML.PREDICT` function.

Task 1. Setup

For each lab, you get a new Google Cloud project and set of resources for a fixed time at no cost.

1. Sign in to Qwiklabs using an **incognito window**.
2. Note the lab's access time (for example, 02:00:00), and make sure you can finish within that time. There is no pause feature. You can restart if needed, but you have to start at the beginning.
3. When ready, click **Start lab**.
4. Note your lab credentials (**Username** and **Password**). You will use them to sign in to the Google Cloud Console.
5. Click **Open Google Console**.
6. Click **Use another account** and copy/paste credentials for **this** lab into the prompts. If you use other credentials, you'll receive errors or **incur charges**.
7. Accept the terms and skip the recovery resource page.

Do not click **End Lab** unless you have finished the lab or want to restart it. This clears your work and removes the project.

Enable the BigQuery API

1. In the Cloud Console, on the **Navigation menu** (), click **APIs & services > Library**.

2. Search for **BigQuery API**, and then click **Enable** if it isn't already enabled.

Task 2. Create your dataset

The first step is to create a BigQuery dataset to store your ML model. To create your dataset:

1. In the Cloud Console, on the **Navigation menu**, click **BigQuery**.

2. In the **Explorer** panel, click the **View actions** icon (three vertical dots) next to your project ID, and then click **Create dataset**.

3. On the Create dataset page:

- For **Dataset ID**, type **bqml_tutorial**

- (Optional) For **Data location**, select **us (multiple regions in United States)**. Currently, the public datasets are stored in the US multi-region location. For simplicity, you should place your dataset in the same location.

4. Leave the remaining settings as their defaults, and click **Create Dataset**.

Task 3. Create your model

Next, you create a linear regression model using the penguins table for BigQuery. The following standard SQL query is used to create the model you use to predict the weight of a penguin:

```
#standardSQL
CREATE OR REPLACE MODEL `bqml_tutorial.penguins_model`
OPTIONS
  (model_type='linear_reg',
  input_label_cols=['body_mass_g']) AS
SELECT
  *
FROM
  `bigquery-public-data.ml_datasets.penguins`
WHERE
  body_mass_g IS NOT NULL
```

In addition to creating the model, running the `CREATE MODEL` command trains the model you create.

Query details

The `CREATE MODEL` clause is used to create and train the model named `bqml_tutorial.penguins_model`.
The `OPTIONS(model_type='linear_reg', input_label_cols=['body_mass_g'])` clause indicates that you are creating a linear regression model. A linear regression is a type of regression model that generates a continuous value from a linear combination of input features.
The `body_mass_g` column is the input label column. For linear regression models, the label column must be real-valued (the column values must be real numbers).
This query's `SELECT` statement uses all the columns in the `bigquery-public-data.ml_datasets.penguins` table. This table contains the following columns that will all be used to predict a penguin's weight:

- `species`: Species of penguin (STRING)
- `island`: Island that the penguin lives on (STRING)
- `culmen_length_mm`: Length of culmen in millimeters (FLOAT64)
- `culmen_depth_mm`: Depth of culmen in millimeters (FLOAT64)
- `flipper_length_mm`: Length of the flipper in millimeters (FLOAT64)
- `sex`: The sex of the penguin (STRING)

The `FROM` clause — `bigquery-public-data.ml_datasets.penguins` — indicates that you are querying the penguins table in the `ml_datasets` dataset. This dataset is in the `bigquery-public-data` project.

The `WHERE` clause — `WHERE body_mass_g IS NOT NULL` — excludes rows where `body_mass_g` is NULL.

Run the CREATE MODEL query

To run the `CREATE MODEL` query to create and train your model:

1. In the Cloud Console, click **Compose new query**.

2. In the **Query editor** text area, enter the following standard SQL query:

```
#standardSQL
CREATE OR REPLACE MODEL `bqml_tutorial.penguins_model`
OPTIONS
  (model_type='linear_reg',
    input_label_cols=['body_mass_g']) AS
SELECT
  *
FROM
  `bigquery-public-data.ml_datasets.penguins`
WHERE
  body_mass_g IS NOT NULL
```

Copied!
content_copy

3. Click **Run**.

The query takes about 30 seconds to complete, after which your model (`penguins_model`) appears in the navigation panel. Because the query uses a `CREATE MODEL` statement to create a table, you do not see query results.

Note
You can ignore the warning about NULL values for input data.

Task 4. Get training statistics (Optional)

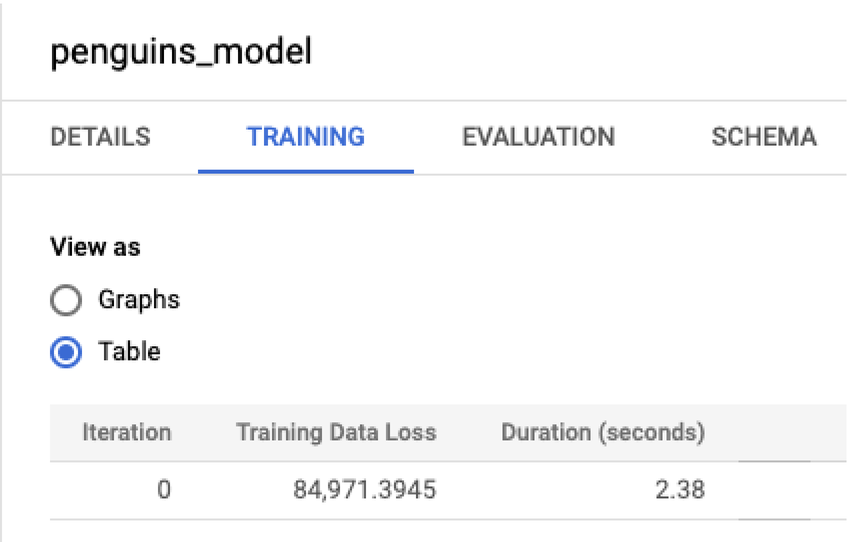
To see the results of the model training, you can use the `ML.TRAINING_INFO` function, or you can view the statistics in the Cloud Console. In this tutorial, you use the Cloud Console.
A machine learning algorithm builds a model by examining many examples and attempting to find a model that minimizes loss. This process is called *empirical risk minimization*.

Loss is the penalty for a bad prediction: a number indicating how bad the model's prediction was on a single example. If the model's prediction is perfect, the loss is zero; otherwise, the loss is greater. The goal of training a model is to find a set of weights and biases that have low loss, on average, across all examples.

To see the model training statistics that were generated when you ran the `CREATE MODEL` query:

1. In the Cloud Console navigation panel, in the **Explorer** section, expand **[PROJECT_ID] > bqml_tutorial > Models (1)**, and then click **penguins_model**.

2. Click the **Training** tab, and then click **Table**. The results should look like the following:



The **Training Data Loss** column represents the loss metric calculated after the model is trained on the training dataset. Because you performed a linear regression, this column is the mean squared error. A "normal_equation" optimization strategy is automatically used for this training, so only one iteration is required to converge to the final model. For more details on the `optimize_strategy` option, see the CREATE MODEL statement for generalized linear models.
For more details on the `ML.TRAINING_INFO` function and "optimize_strategy" training option, see the BigQuery ML syntax reference.

Task 5. Evaluate your model

After creating your model, you evaluate the performance of the model using the ML.EVALUATE function. The ML.EVALUATE function evaluates the predicted values against the actual data.

The following query is used to evaluate the model:

```
#standardSQL
SELECT
  *
FROM
  ML.EVALUATE(MODEL `bqml_tutorial.penguins_model`,
    (
    SELECT
      *
    FROM
      `bigquery-public-data.ml_datasets.penguins`
    WHERE
      body_mass_g IS NOT NULL))
```

Query details

The first SELECT statement retrieves the columns from your model.

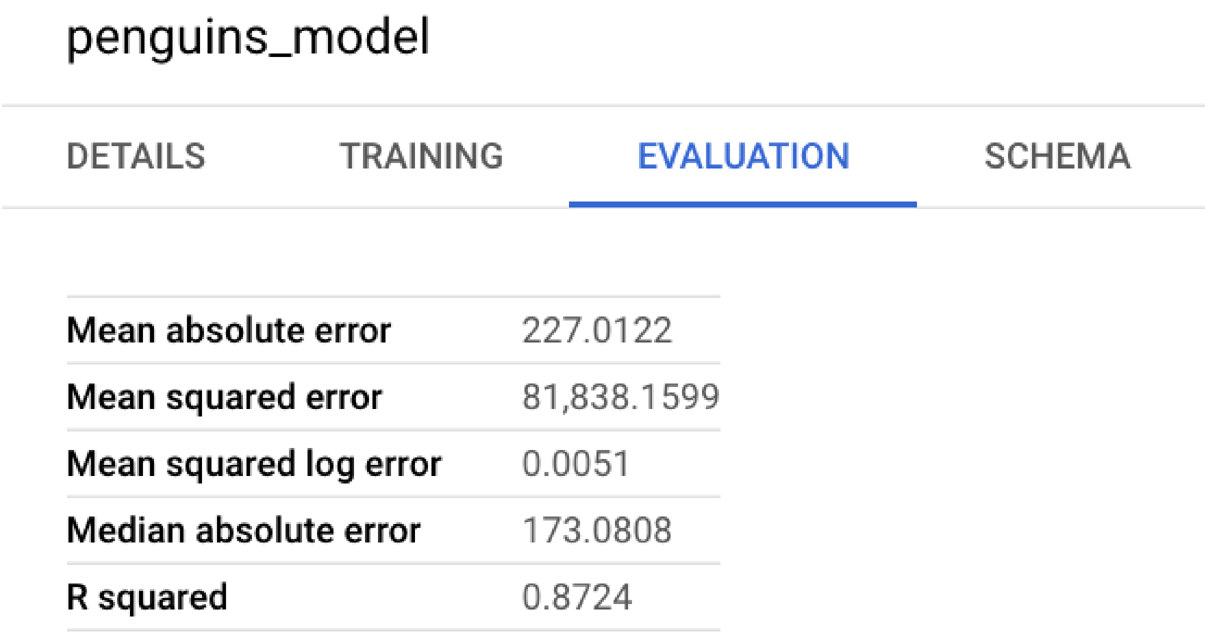
The FROM clause uses the ML.EVALUATE function against your model: bqml_tutorial.penguins_model.
This query's nested SELECT statement and FROM clause are the same as those in the CREATE MODEL query.

The WHERE clause — WHERE body_mass_g IS NOT NULL — excludes rows where body_mass_g is NULL.

A proper evaluation would be on a subset of the penguins table that is separate from the data used to train the model. You can also call ML.EVALUATE without providing the input data. ML.EVALUATE will retrieve the evaluation metrics calculated during training, which uses the automatically reserved evalution dataset:

```
#standardSQL
SELECT
  *
FROM
  ML.EVALUATE(MODEL `bqml_tutorial.penguins_model`)
```

You can also use the Cloud Console to view the evaluation metrics calculated during the training. The results should look like the following:



Mean absolute error	227.0122
Mean squared error	81,838.1599
Mean squared log error	0.0051
Median absolute error	173.0808
R squared	0.8724

Run the ML.EVALUATE query

To run the `ML.EVALUATE` query that evaluates the model:

1. In the Cloud Console, click **Compose new query**.

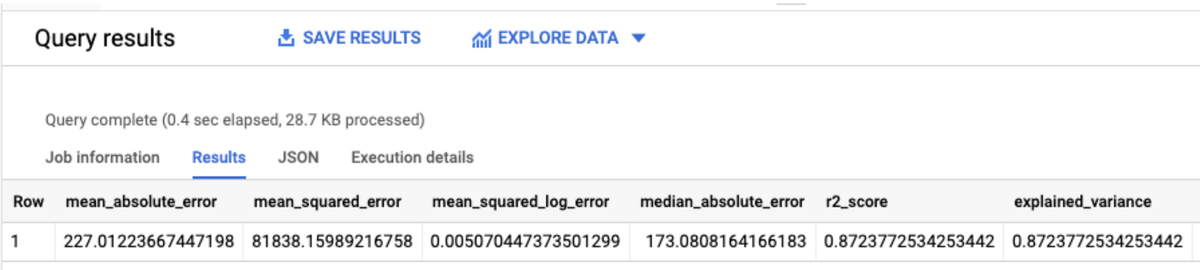
2. In the **Query editor** text area, enter the following standard SQL query:

```
#standardSQL
SELECT
  *
FROM
  ML.EVALUATE(MODEL `bqml_tutorial.penguins_model`,
    (
    SELECT
      *
    FROM
      `bigquery-public-data.ml_datasets.penguins`
    WHERE
      body_mass_g IS NOT NULL))
```
Copied!
content_copy

3. (Optional) To set the data location, click **More > Query settings**. For **Data location**, select **us (multiple regions in United States)**.

4. Click **Run**.

5. When the query is complete, click the **Results** tab below the query text area. The results should look like the following:



Because you performed a linear regression, the results include the following columns:

- `mean_absolute_error`
- `mean_squared_error`
- `mean_squared_log_error`
- `median_absolute_error`
- `r2_score`
- `explained_variance`

An important metric in the evaluation results is the [R^2] score. The R^2 score is a statistical measure that determines whether the linear regression predictions approximate the actual data. 0 indicates that the model explains none of the variability of the response data around the mean. 1 indicates that the model explains all the variability of the response data around the mean.

Task 6. Use your model to predict outcomes

Now that you have evaluated your model, the next step is to use it to predict an outcome. You use your model to predict the body mass in grams of all penguins that reside in Biscoe.

The following query is used to predict the outcome:

```
#standardSQL
SELECT
  *
FROM
```

```
  ML.PREDICT(MODEL `bqml_tutorial.penguins_model`,
    (
    SELECT
      *
    FROM
      `bigquery-public-data.ml_datasets.penguins`
    WHERE
      body_mass_g IS NOT NULL
      AND island = "Biscoe"))
```

Query details

The first `SELECT` statement retrieves the `predicted_body_mass_g` column along with the columns in `bigquery-public-data.ml_datasets.penguins`. This column is generated by the `ML.PREDICT` function. When you use the `ML.PREDICT` function, the output column name for the model is `predicted_<label_column_name>`. For linear regression models, `predicted_label` is the estimated value of `label`. For logistic regression models, `predicted_label` is one of the two input labels depending on which label has the higher predicted probability.

The `ML.PREDICT` function is used to predict results using your model: `bqml_tutorial.penguins_model`.
This query's nested SELECT `statement` and `FROM` clause are the same as those in the `CREATE MODEL` query.

The `WHERE` clause — `WHERE island = "Biscoe"` — indicates that you are limiting the prediction to the island of Biscoe.

Run the ML.PREDICT query

To run the query that uses the model to predict an outcome:

1. In the Cloud Console, click **Compose new query**.
2. In the **Query editor** text area, enter the following standard SQL query:

```
#standardSQL
  SELECT
    *
  FROM
    ML.PREDICT(MODEL `bqml_tutorial.penguins_model`,
      (
```

```
SELECT
  *
FROM
  `bigquery-public-data.ml_datasets.penguins`
WHERE
  body_mass_g IS NOT NULL
  AND island = "Biscoe"))
```
Copied!
content_copy

3. (Optional) To set the data location, click **More > Query settings**. For **Data location**, select **us (multiple regions in United States)**.

4. Click **Run**.

5. When the query is complete, click the **Results** tab below the query text area. The results should look like the following:

Query results SAVE RESULTS EXPLORE DATA

Query complete (0.3 sec elapsed, 28.6 KB processed)

Job information Results JSON Execution details

Row	predicted_body_mass_g	species	island	culmen_length_mm	culmen_depth_mm	flipper_length_mm	body_mass_g	sex
1	3875.1671265197947	Adelie Penguin (Pygoscelis adeliae)	Biscoe	39.7	18.9	184.0	3550.0	MALE
2	3365.9310542364647	Adelie Penguin (Pygoscelis adeliae)	Biscoe	36.4	17.1	184.0	2850.0	FEMALE
3	4063.638353343009	Adelie Penguin (Pygoscelis adeliae)	Biscoe	41.6	18.0	192.0	3950.0	MALE
4	3529.278475013224	Adelie Penguin (Pygoscelis adeliae)	Biscoe	35.0	17.9	192.0	3725.0	FEMALE
5	4058.1285239777285	Adelie Penguin (Pygoscelis adeliae)	Biscoe	41.1	18.2	192.0	4050.0	MALE
6	4288.827255164885	Adelie Penguin (Pygoscelis adeliae)	Biscoe	42.0	19.5	200.0	4050.0	MALE
7	4538.440797625522	Gentoo penguin (Pygoscelis papua)	Biscoe	43.8	13.9	208.0	4300.0	FEMALE
8	4529.972792532769	Gentoo penguin (Pygoscelis papua)	Biscoe	43.3	14.0	208.0	4575.0	FEMALE
9	4534.136742771194	Gentoo penguin (Pygoscelis papua)	Biscoe	44.0	13.6	208.0	4350.0	FEMALE
10	4507.386848366082	Gentoo penguin (Pygoscelis papua)	Biscoe	42.7	13.7	208.0	3950.0	FEMALE
11	4569.761164358724	Gentoo penguin (Pygoscelis papua)	Biscoe	45.3	13.8	208.0	4200.0	FEMALE

Task 7. Explain prediction results with explainable AI methods

To understand why your model is generating these prediction results, you can use the `ML.EXPLAIN_PREDICT` function.

`ML.EXPLAIN_PREDICT` is an extended version of `ML.PREDICT`. `ML.EXPLAIN_PREDICT` returns prediction results with additional columns that explain those results. You can run `ML.EXPLAIN_PREDICT` without `ML.PREDICT`. For an in-depth explanation of Shapley values and explainable AI in BigQuery ML, see [BigQuery ML explainable AI overview].
The following query is used to generate explanations:

```
#standardSQL
SELECT
  *
FROM
  ML.EXPLAIN_PREDICT(MODEL `bqml_tutorial.penguins_model`,
    (
    SELECT
      *
    FROM
      `bigquery-public-data.ml_datasets.penguins`
    WHERE
      body_mass_g IS NOT NULL
      AND island = "Biscoe"),
    STRUCT(3 as top_k_features))
```

Query details

Run the ML.EXPLAIN_PREDICT query

To run the `ML.EXPLAIN_PREDICT` query that explains the model:

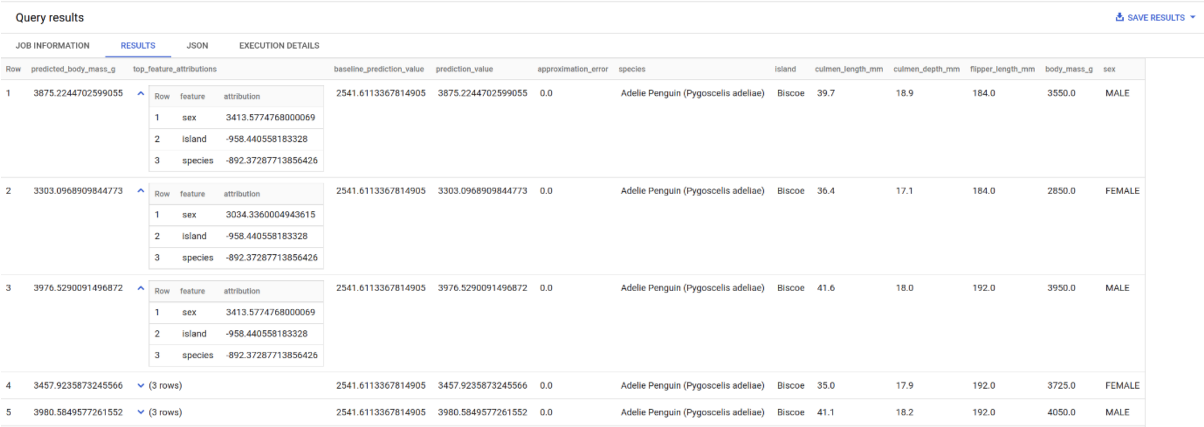
1. In the Cloud Console, click **Compose new query**.

2. In the **Query editor** text area, enter the following standard SQL query:

```
#standardSQL
SELECT
*
FROM
ML.EXPLAIN_PREDICT(MODEL `bqml_tutorial.penguins_model`,
  (
  SELECT
    *
  FROM
    `bigquery-public-data.ml_datasets.penguins`
  WHERE
    body_mass_g IS NOT NULL
    AND island = "Biscoe"),
  STRUCT(3 as top_k_features))
```
Copied!
content_copy

3. Click **Run**.

4. When the query is complete, click the **Results** tab below the query text area. The results should look like the following:



Query results

JOB INFORMATION | RESULTS | JSON | EXECUTION DETAILS | SAVE RESULTS

Row	predicted_body_mass_g	top_feature_attributions	baseline_prediction_value	prediction_value	approximation_error	species	island	culmen_length_mm	culmen_depth_mm	flipper_length_mm	body_mass_g	sex
1	3875.2244702599055	Row feature attribution 1 sex 3413.5774768000069 2 island -958.440558183328 3 species -892.37287713856426	2541.6113367814905	3875.2244702599055	0.0	Adelie Penguin (Pygoscelis adeliae)	Biscoe	39.7	18.9	184.0	3550.0	MALE
2	3303.0968909844773	Row feature attribution 1 sex 3034.3360004943615 2 island -958.440558183328 3 species -892.37287713856426	2541.6113367814905	3303.0968909844773	0.0	Adelie Penguin (Pygoscelis adeliae)	Biscoe	36.4	17.1	184.0	2850.0	FEMALE
3	3976.5290091496872	Row feature attribution 1 sex 3413.5774768000069 2 island -958.440558183328 3 species -892.37287713856426	2541.6113367814905	3976.5290091496872	0.0	Adelie Penguin (Pygoscelis adeliae)	Biscoe	41.6	18.0	192.0	3950.0	MALE
4	3457.9235873245566	(3 rows)	2541.6113367814905	3457.9235873245566	0.0	Adelie Penguin (Pygoscelis adeliae)	Biscoe	35.0	17.9	192.0	3725.0	FEMALE
5	3980.5849577261552	(3 rows)	2541.6113367814905	3980.5849577261552	0.0	Adelie Penguin (Pygoscelis adeliae)	Biscoe	41.1	18.2	192.0	4050.0	MALE

Note The **ML.EXPLAIN_PREDICT** query outputs all the input feature columns, similar to what **ML.PREDICT** does. Only one feature column, "species", is shown in the figure above for readability purposes.
For linear regression models, Shapley values are used to generate feature attribution values per feature in the model. `ML.EXPLAIN_PREDICT` outputs the top 3 feature attributions per row of the table provided because `top_k_features` was set to 3 in the query. These attributions are sorted by the absolute value of the attribution in descending order. In all examples, the feature `sex` contributed the most to the overall prediction. For detailed explanations of the output columns of the `ML.EXPLAIN_PREDICT` query, see `ML.EXPLAIN_PREDICT` syntax documentation

Task 8. Globally explain your model (Optional)

To know which features are the most important to determine the weights of the penguins in general, you can use the `ML.GLOBAL_EXPLAIN` function. In order to use `ML.GLOBAL_EXPLAIN`, the model must be retrained with the option `ENABLE_GLOBAL_EXPLAIN=TRUE`. Rerun the training query with this option using the following query:

```
#standardSQL
CREATE OR REPLACE MODEL bqml_tutorial.penguins_model
```

```
OPTIONS
  (model_type='linear_reg',
  input_label_cols=['body_mass_g'],
  enable_global_explain=TRUE) AS
SELECT
  *
FROM
  `bigquery-public-data.ml_datasets.penguins`
WHERE
  body_mass_g IS NOT NULL
```
Copied!
content_copy
Note
You can ignore the warning about NULL values for input data.

Access global explanations through ML.GLOBAL_EXPLAIN

The following query is used to generate global explanations:

```
#standardSQL
SELECT
  *
FROM
  ML.GLOBAL_EXPLAIN(MODEL `bqml_tutorial.penguins_model`)
```

Query details

Run the ML.GLOBAL_EXPLAIN query

To run the `ML.GLOBAL_EXPLAIN` query:

1. In the Cloud Console, click **Compose new query**.

2. In the **Query editor** text area, enter the following standard SQL query:

```
#standardSQL
SELECT
*
FROM
ML.GLOBAL_EXPLAIN(MODEL `bqml_tutorial.penguins_model`)
```
Copied!
content_copy

3. (Optional) To set the data location, click **More > Query settings**. For **Data location**, select **us (multiple regions in United States)**.

4. Click **Run**.

5. When the query is complete, click the **Results** tab below the query text area. The results should look like the following:

Job information Results JSON Execution

Row	feature	attribution
1	sex	3036.706115070624
2	species	514.902716582866
3	flipper_length_mm	193.61205101879835
4	culmen_depth_mm	117.0849441916805
5	culmen_length_mm	94.3667930401437
6	island	13.975971094405732

Task 9. Clean up

To avoid incurring charges to your Google Cloud account for the resources used in this tutorial, either delete the project that contains the resources, or keep the project and delete the individual resources.

Deleting your dataset

Deleting your project removes all datasets and all tables in the project. If you prefer to reuse the project, you can delete the dataset you created in this tutorial:

1. If necessary, open the BigQuery page in the Cloud Console.

2. In the **Explorer** panel, click the **View actions** icon (⋮) next to your dataset. Click **Delete**.

3. In the Delete dataset dialog box, to confirm the delete command, type **delete** and then click **Delete**.

Deleting your project

To delete the project:

1. In the Cloud Console, on the **Navigation menu**, click **IAM & Admin > Manage Resources**.

Note If prompted, Click **LEAVE** for unsaved work.

2. In the project list, select the project that you want to delete, and then click **Delete**.

3. In the dialog, type the project ID, and then click **Shut down** to delete the project.

Congratulations!

You've learned how to:

- Create a linear regression model using the `CREATE MODEL` statement with BigQuery ML.

- Evaluate the ML model with the `ML.EVALUATE` function.

- Make predictions using the ML model with the `ML.PREDICT` function.

What's next

- To learn more about machine learning, see the [Machine learning crash course]().
- For an overview of BigQuery ML, see Introduction to [BigQuery ML]().
- To learn more about the Cloud Console, see [Using the Cloud Console]().

End your lab

When you have completed your lab, click **End Lab**. Qwiklabs removes the resources you've used and cleans the account for you.

You will be given an opportunity to rate the lab experience. Select the applicable number of stars, type a comment, and then click **Submit**.

The number of stars indicates the following:

- 1 star = Very dissatisfied
- 2 stars = Dissatisfied
- 3 stars = Neutral
- 4 stars = Satisfied
- 5 stars = Very satisfied

You can close the dialog box if you don't want to provide feedback.

For feedback, suggestions, or corrections, please use the **Support** tab.