15 Basic And Highly Used Hive Queries that All Data Engineers Must know

SIDDHARTH SONKAR (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/AUTHOR/SIDDHARTH_S/), DECEMBER 1, 2020 LOGIN TO BOOKMARK THIS ...

Article   🔒 Video Book
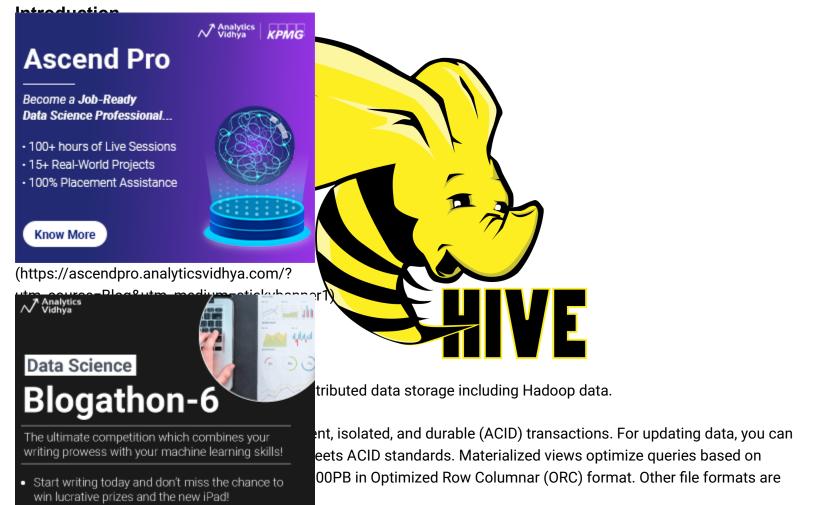
Overview

- Get to know 15 basic hive queries including-
  - Simple selects - selecting columns
  - Simple selects – selecting rows
  - Creating new columns
  - Hive Functions

- By no way is this list exhaustive. Feel free to add more in the comments section.

Introduction

tributed data storage including Hadoop data.

ent, isolated, and durable (ACID) transactions. For updating data, you can
eets ACID standards. Materialized views optimize queries based on
00PB in Optimized Row Columnar (ORC) format. Other file formats are

You can create tables that resemble those in a traditional relational database. You use familiar insert, update, delete, and merge SQL statements to query table data.

The insert statement writes data to tables. Update and delete statements modify and delete values already written to Hive. The merge statement streamlines updates, deletes, and changes data capture operations by drawing on co-existing tables.

These statements support auto-commit that treats each statement as a separate transaction and commits it after the SQL statement is executed. In this article, we are going to cover some basic Hive queries and functions that will give you a basic understanding of how to run Hive queries over distributed data.

Table of Contents:

- Concat
- Substr

CT statement. A select statement has 6 key components;

- FROM table-name
- GROUP BY column names

ese clauses in them simplifying many queries. On the other hand,
complex and if you need to JOIN two or more tables together then more

ten in uppercase for clarity. HQL is not case sensitive. Neither do you
is often clearer to do so for all but the simplest of queries.

In this lesson, we will start with the very simple and work our way up to the more complex.

Simple selects – selecting columns
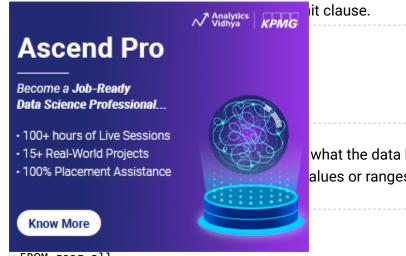
Amongst all the hive queries, the simplest query is effectively one which returns the contents of the whole table

```
SELECT *

FROM geog_all;

CREATE VIEW [IF NOT EXISTS] [db_name.]view_name [(column_name [COMMENT column_comment], ...)  ]
```

It is better to practice, and generally more efficient to explicitly list the column names that you want to be returned.

```
SELECT anonid, fueltypes, acorn_type

FROM geog_all;
```

Simple selects – selecting rows

In addition to limiting the columns returned by a query, you can also limit the rows returned. The simplest case is to ...it clause.

what the data looks like. Usually, you will want to restrict the rows ...alues or ranges within one or more columns.

FROM geog_all]

ore complex and involve more than one column.

ype > 42;

WHERE fueltypes = "ElecOnly" AND acorn_type > 42 AND nuts1 <> "--";

Notice that the columns used in the conditions of the Where clause don't have to appear in the Select clause. Other operators can also be used in the where clause. For complex expressions, brackets can be used to enforce precedence.

```
SELECT anonid, fueltypes, acorn_type, nuts1, ldz

FROM geog_all

WHERE

fueltypes = "ElecOnly"

AND acorn_type BETWEEN 42 AND 47

AND (nuts1 NOT IN ("UKM", "UKI") OR ldz = "--");
```

Creating new columns
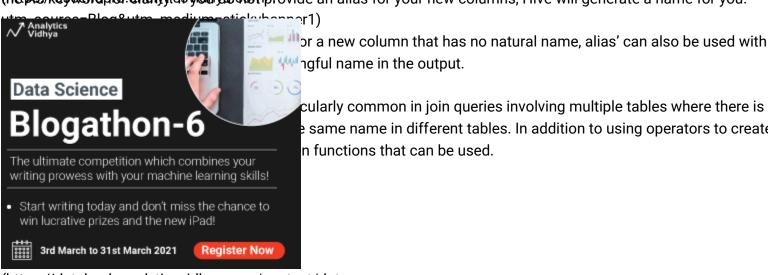
It is possible to create new columns in the output of the query. These columns can be from combinations from the other columns using operators and/or built-in Hive functions.

(eprofileclass * acorn_type) AS multiply, (eprofileclass + acorn_

...able within the Hive can be found in the [documentation](.../Hive/LanguageManual+UDF).

...provide an 'alias' for the column. This is essentially the name you wish to ...mediately after the expression to which it refers. Optionally you can add

...provide an alias for your new columns, Hive will generate a name for you.

...or a new column that has no natural name, alias' can also be used with ...ngful name in the output.

...cularly common in join queries involving multiple tables where there is a ...e same name in different tables. In addition to using operators to create ...n functions that can be used.

Simple functions

Concat can be used to add strings together.
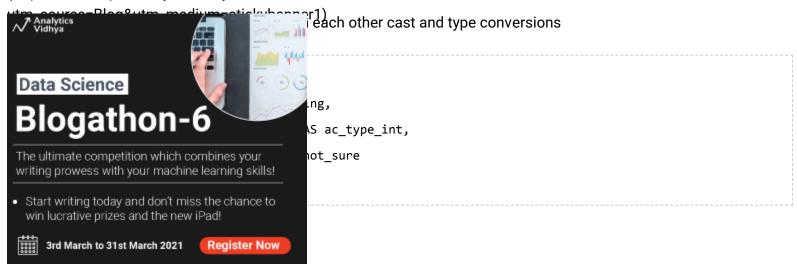
```
SELECT anonid, acorn_category,

acorn_group,

acorn_type,

concat (acorn_category, ",", acorn_group, ",", acorn_type)  AS acorn_code

FROM geog_all;
```

substr can be used to extract a part of a string

```
SELECT anon_id,

advancedatetime,

substr (advancedatetime, 1, 2) AS day,

substr (advancedatetime, 3, 3) AS month,

substr (advancedatetime, 6, 2) AS year
```

```
          substr (advancedtime, 6, 2) AS year
   FROM elec_c;
```

... each other cast and type conversions

```
                              AS reverse_a_typepo
```

```
                              ing,
                              AS ac_type_int,
                              not_sure
```

Aggregations

Aggregate functions are used to perform some kind of mathematical or statistical calculation across a group of rows. The rows in each group are determined by the different values in a specified column or columns. A list of all of the available functions is available in the apache documentation.

```
SELECT anon_id,
          count (eleckwh) AS total_row_count,
          sum (eleckwh) AS total_period_usage,
          min (eleckwh) AS min_period_usage,
          avg (eleckwh) AS avg_period_usage,
          max (eleckwh) AS max_period_usage
       FROM elec_c
GROUP BY anon_id;
```
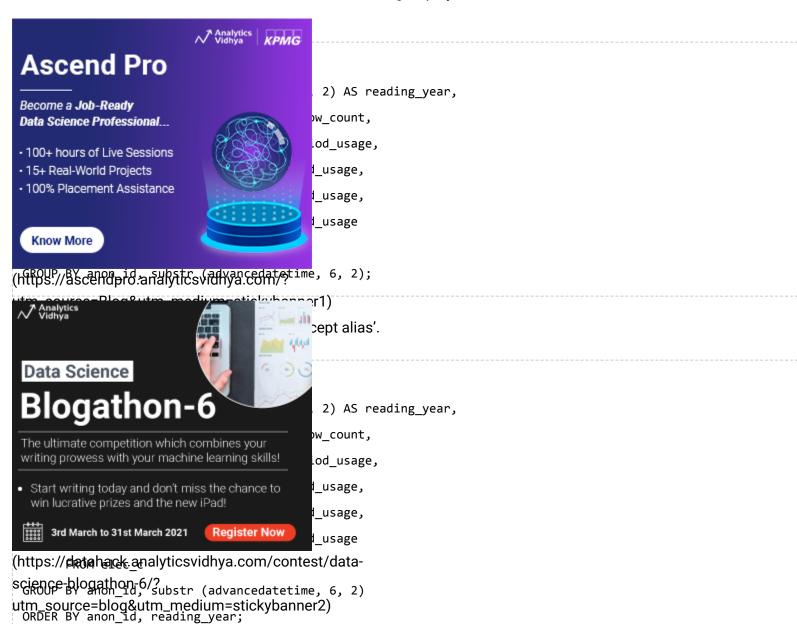
In the above example, five aggregations were performed over the single column anon_id. It is possible to aggregate over multiple columns by specifying them in both the select and the group by clause. The grouping will take place

based on the order of the columns listed in the group by clause. What is not allowed is specifying a non-aggregated column in the select clause which is not mentioned in the group by clause.

```
           2) AS reading_year,
      ow_count,
      lod_usage,
      d_usage,
      d_usage,
      d_usage

GROUP BY anon_id, substr (advancedatetime, 6, 2);
```

cept alias'.

```
           2) AS reading_year,
      ow_count,
      lod_usage,
      d_usage,
      d_usage,
      d_usage
FROM elec_c
GROUP BY anon_id, substr (advancedatetime, 6, 2)
ORDER BY anon_id, reading_year;
```

But the Order by clause does.
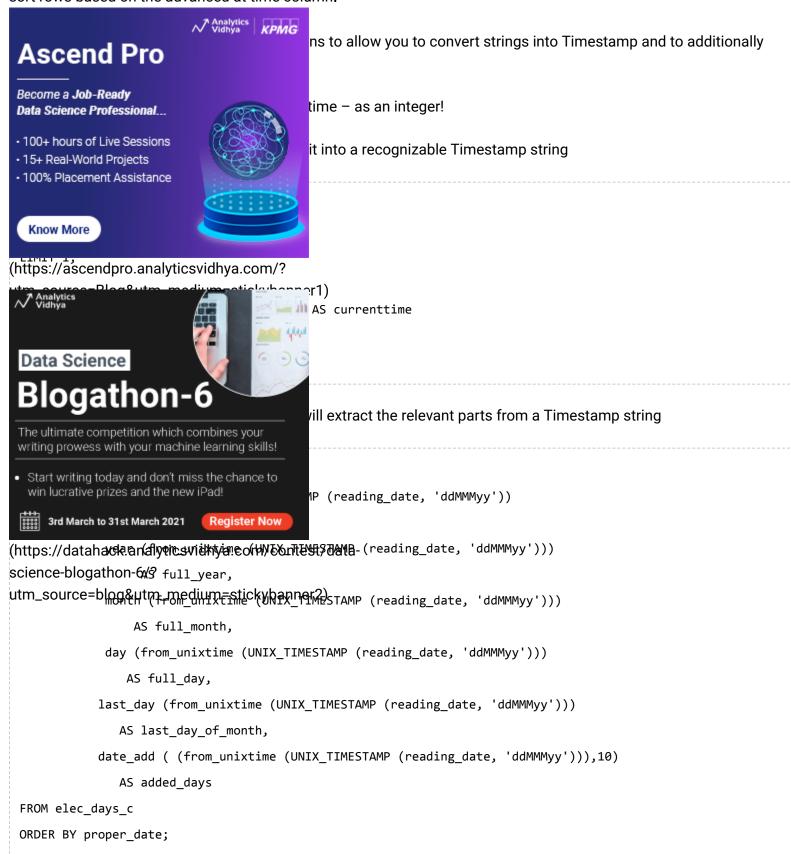
The Distinct keyword provides a set of a unique combination of column values within a table without any kind of aggregation.

```
SELECT DISTINCT eprofileclass, fueltypes

FROM geog_all;
```

Date functions
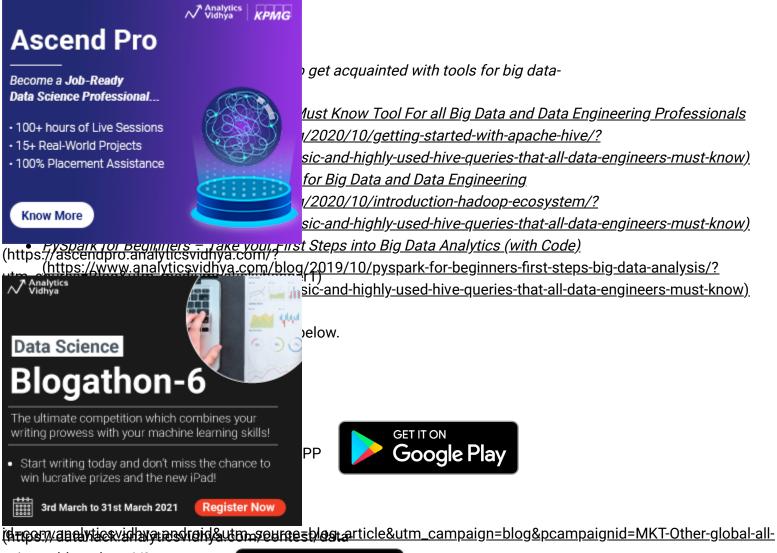
In the elec_c and gas_c tables, the advance DateTime column, although it contains timestamp type information, it is
defined as a string type. For much of the time, this can be quite convenient. However, there will be times when we

defined as a string type. For much of the time, this can be quite convenient. However, there will be times when we really do need to be able to treat the column as a Timestamp. Perhaps the most obvious example is when you need to sort rows based on the advanced at time column.

...ns to allow you to convert strings into Timestamp and to additionally

...time – as an integer!

...it into a recognizable Timestamp string

... AS currenttime

...will extract the relevant parts from a Timestamp string

```
...MP (reading_date, 'ddMMMyy'))

year (from_unixtime (UNIX_TIMESTAMP (reading_date, 'ddMMMyy')))
            AS full_year,
      month (from_unixtime (UNIX_TIMESTAMP (reading_date, 'ddMMMyy')))
            AS full_month,
      day (from_unixtime (UNIX_TIMESTAMP (reading_date, 'ddMMMyy')))
            AS full_day,
      last_day (from_unixtime (UNIX_TIMESTAMP (reading_date, 'ddMMMyy')))
            AS last_day_of_month,
      date_add ( (from_unixtime (UNIX_TIMESTAMP (reading_date, 'ddMMMyy'))),10)
            AS added_days
FROM elec_days_c
ORDER BY proper_date;
```

Conclusion:

In the article, we covered some basic Hive functions and queries. Running queries on distributed data is not much different from running queries in MySQL. This will follow a few articles in which we will cover more advanced functions and queries. I hope you might have enjoyed the article. Please don't forget to drop in your comments in the comments

get acquainted with tools for big data-

Must Know Tool For all Big Data and Data Engineering Professionals
/2020/10/getting-started-with-apache-hive/?
sic-and-highly-used-hive-queries-that-all-data-engineers-must-know)
for Big Data and Data Engineering
/2020/10/introduction-hadoop-ecosystem/?
sic-and-highly-used-hive-queries-that-all-data-engineers-must-know)
PySpark for Beginners – Take your First Steps into Big Data Analytics (with Code)

(https://www.analyticsvidhya.com/blog/2019/10/pyspark-for-beginners-first-steps-big-data-analysis/?
sic-and-highly-used-hive-queries-that-all-data-engineers-must-know)

below.

GET IT ON
Google Play

id=com.analyticsvidhya.android&utm_source=blog_article&utm_campaign=blog&pcampaignid=MKT-Other-global-all-

DOWNLOAD ON THE
App Store   (https://apps.apple.com/us/app/analytics-

vidhya/id1470025572)

Related Articles
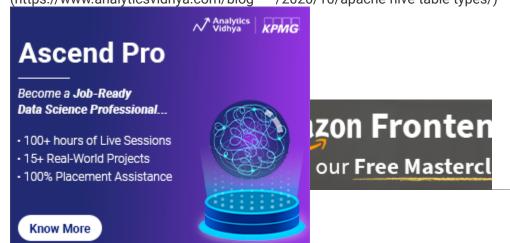
(https://www.analyticsvidhya.com/blog/2020/10/getting-
started-with-apache-hive/)
Getting Started with Apache Hive - A

(https://www.analyticsvidhya.com/blog/2020/10/apache-
hive-table-types/)
Types of Tables in Apache Hive - A

(https://www.analyticsvidhya.com/blog/2020/10/apache-
sql-rdbms/)
Basics of SQL and RDBMS - must have

Must Know Tool For all Big Data and Data Engineering Professionals (https://www.analyticsvidhya.com/blog

Quick Overview (https://www.analyticsvidhya.com/blog/2020/10/apache-hive-table-types/)

skills for data science professionals (https://www.analyticsvidhya.com/blog/2015/03/basics-sql-rdbms/)

TAGS : CREATE NEW ROWS (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/TAG/CREATE-NEW-ROWS/), SELECT COLUMNS (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/TAG/SELECT-COLUMNS/), SELECT ROWS (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/TAG/SELECT-ROWS/)

NEXT ARTICLE

Use of Machine Learning in Dairy farming
(https://www.analyticsvidhya.com/blog/2020/12/use-of-machine-learning-in-dairy-farming/)
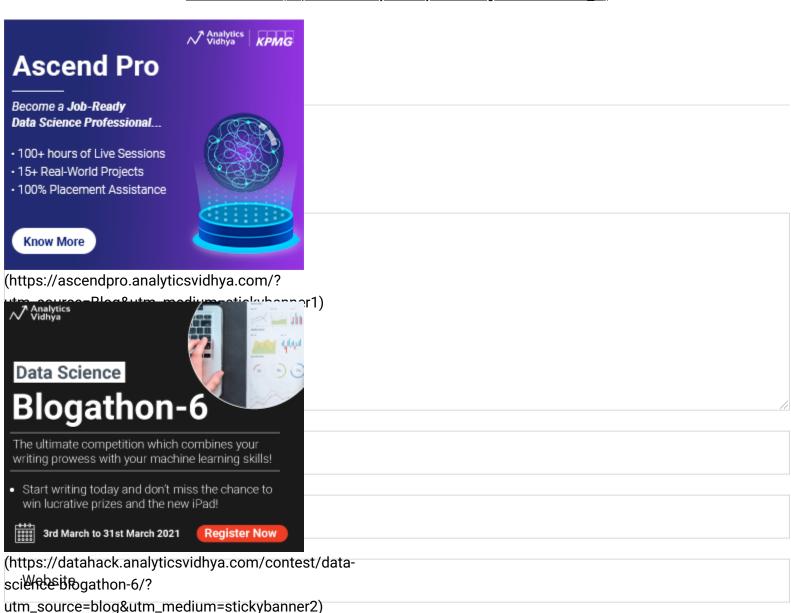
• • •

PREVIOUS ARTICLE

A Case Study To Detect Anomalies In Time Series Using Anomalize Package In R

(https://www.analyticsvidhya.com/blog/2020/12/a-case-study-to-detect-anomalies-in-time-series-using-anomalize-package-in-r/)

Website
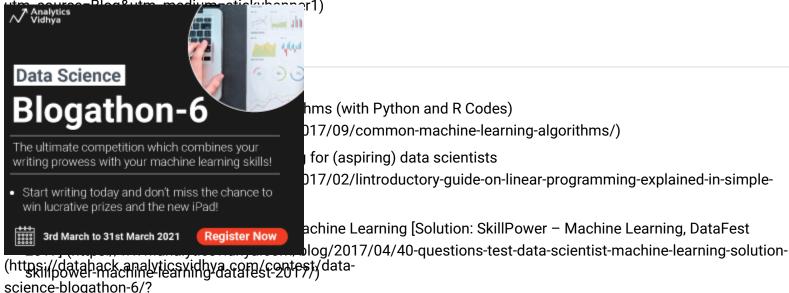
☐ Notify me of new posts by email.

SUBMIT COMMENT

hms (with Python and R Codes)
017/09/common-machine-learning-algorithms/)

for (aspiring) data scientists
017/02/lintroductory-guide-on-linear-programming-explained-in-simple-

achine Learning [Solution: SkillPower – Machine Learning, DataFest
blog/2017/04/40-questions-test-data-scientist-machine-learning-solution-skillpower-machine-learning-datafest-2017/)

6 Easy Steps to Learn Naive Bayes Algorithm with codes in Python and R
(https://www.analyticsvidhya.com/blog/2017/09/naive-bayes-explained/)

Customer Sentiments Analysis of Pepsi and Coca-Cola using Twitter Data in R
(https://www.analyticsvidhya.com/blog/2021/02/customer-sentiments-analysis-of-pepsi-and-coca-cola-using-
twitter-data-in-r/)

45 Questions to test a data scientist on basics of Deep Learning (along with solution)
(https://www.analyticsvidhya.com/blog/2017/01/must-know-questions-deep-learning/)

40 Questions to test a Data Scientist on Clustering Techniques (Skill test Solution)
(https://www.analyticsvidhya.com/blog/2017/02/test-data-scientist-clustering/)

30 Questions to test a data scientist on Linear Regression [Solution: Skilltest – Linear Regression]
(https://www.analyticsvidhya.com/blog/2017/07/30-questions-to-test-a-data-scientist-on-linear-regression/)

(https://ascendpro.analyticsvidhya.com/?
CAREER RESOURCES
utm_source=blog&utm_medium=stickybanner1)

(https://datahack.analyticsvidhya.com/contest/data-
science-blogathon-6/?
utm_source=blog&utm_medium=stickybanner2)

Key Questions You Should Answer Before Transitioning into Data
Science (https://www.analyticsvidhya.com/16-key-questions-data-science-
career-transition/?
utm_source=Blog&utm_medium=CareerResourceWidget)
NOVEMBER 23, 2020

Here's What You Need to Know to Become a Data Scientist!
(https://www.analyticsvidhya.com/blog/2021/01/heres-what-you-need-to-
know-to-become-a-data-scientist/?
&utm_source=Blog&utm_medium=CareerResourceWidget)
JANUARY 22, 2021

These 7 Signs Show you have Data Scientist Potential!
(https://www.analyticsvidhya.com/blog/2020/12/these-7-signs-show-you-
have-data-scientist-potential/?
&utm_source=Blog&utm_medium=CareerResourceWidget)
DECEMBER 3, 2020

How To Have a Career in Data Science (Business Analytics)?
(https://www.analyticsvidhya.com/blog/2020/11/how-to-have-a-career-in-
data-science-business-analytics/?
&utm_source=Blog&utm_medium=CareerResourceWidget)
NOVEMBER 26, 2020

Should I become a data scientist (or a business analyst)?
(https://www.analyticsvidhya.com/blog/2020/11/become-data-scientist-
business-analyst/?
&utm_source=Blog&utm_medium=CareerResourceWidget)

(https://ascendpro.analyticsvidhya.com/?
utm_source=blog&utm_medium=stickybanner1)

ttps://www.analyticsvidhya.com/blog/2021/03/create-your-first-
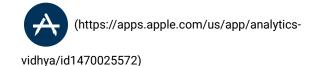
ttps://www.analyticsvidhya.com/blog/2021/03/variants-of-gradient-

Quick Guide To Estimation (https://www.analyticsvidhya.com/blog/2021/03/quick-guide-to-estimation/)

s://www.analyticsvidhya.com/blog/2021/03/understanding-robotics-

(https://datahack.analyticsvidhya.com/contest/data-
science-blogathon-6/?
utm_source=blog&utm_medium=stickybanner2)

(https://play.google.com/store/apps/details?
id=com.analyticsvidhya.android)

(https://apps.apple.com/us/app/analytics-
vidhya/id1470025572)

Analytics Vidhya

About Us (https://www.analyticsvidhya.com/about-me/)

Our Team (https://www.analyticsvidhya.com/about-me/team/)

Careers (https://www.analyticsvidhya.com/about-me/career-analytics-
vidhya/)

Contact us (https://www.analyticsvidhya.com/contact/)

Data Science

Blog (https://www.analyticsvidhya.com/blog/)

Hackathon (https://datahack.analyticsvidhya.com/)

Discussions (https://discuss.analyticsvidhya.com/)

Apply Jobs (https://www.analyticsvidhya.com/jobs/)

Companies

Post Jobs (https://www.analyticsvidhya.com/corporate/)

Trainings (https://courses.analyticsvidhya.com/)

Hiring Hackathons (https://datahack.analyticsvidhya.com/)

Advertising (https://www.analyticsvidhya.com/contact/)

Visit us

(https://www.linkedin.com/company/analytics-
vidhya/)
(https://facebook.com/AnalyticsVidhya)
(https://twitter.com/analyticsvidhya)
(https://www.youtube.com/channel/UC8gDteHtH4hg3o2343iObA)

© Copyright 2013-2020 Analytics Vidhya

Privacy Policy   Terms of Use   Refund Policy