# GSP327: Engineer Data in Google Cloud

Task - 1: Clean your training data

1CREATE OR REPLACE TABLE

2 taxirides.taxi\_training\_data AS

3SELECT

4 (tolls\_amount + fare\_amount) AS fare\_amount,

5 pickup\_datetime,

6 pickup\_longitude AS pickuplon,

7 pickup\_latitude AS pickuplat,

8 dropoff\_longitude AS dropofflon,

9 dropoff\_latitude AS dropofflat,

10 passenger\_count AS passengers,

11FROM

12 taxirides.historical\_taxi\_rides\_raw

13WHERE

14 RAND() < 0.001

15 AND trip\_distance > 0

16 AND fare\_amount >= 2.5

17 AND pickup\_longitude > -78

18 AND pickup\_longitude < -70

19 AND dropoff\_longitude > -78

20 AND dropoff\_longitude < -70

21 AND pickup\_latitude > 37

22 AND pickup\_latitude < 45

23 AND dropoff\_latitude > 37

24 AND dropoff\_latitude < 45

25 AND passenger\_count > 0

### Task - 2: Create a BQML model called taxirides.fare\_model

1CREATE OR REPLACE MODEL taxirides.fare\_model

2TRANSFORM(

3 \* EXCEPT(pickup\_datetime)

4

5 , ST\_Distance(ST\_GeogPoint(pickuplon, pickuplat), ST\_GeogPoint(dropofflon, dropofflat)) AS euclidean

6 , CAST(EXTRACT(DAYOFWEEK FROM pickup\_datetime) AS STRING) AS dayofweek

7 , CAST(EXTRACT(HOUR FROM pickup\_datetime) AS STRING) AS hourofday

8)

9OPTIONS(input\_label\_cols=['fare\_amount'], model\_type='linear\_reg')

10AS

11

12SELECT \* FROM taxirides.taxi\_training\_data

### Task - 3: Perform a batch prediction on new data

1CREATE OR REPLACE TABLE taxirides.2015\_fare\_amount\_predictions

2 AS

3SELECT \* FROM ML.PREDICT(MODEL taxirides.fare\_model,(

4 SELECT \* FROM taxirides.report\_prediction\_data)

5)