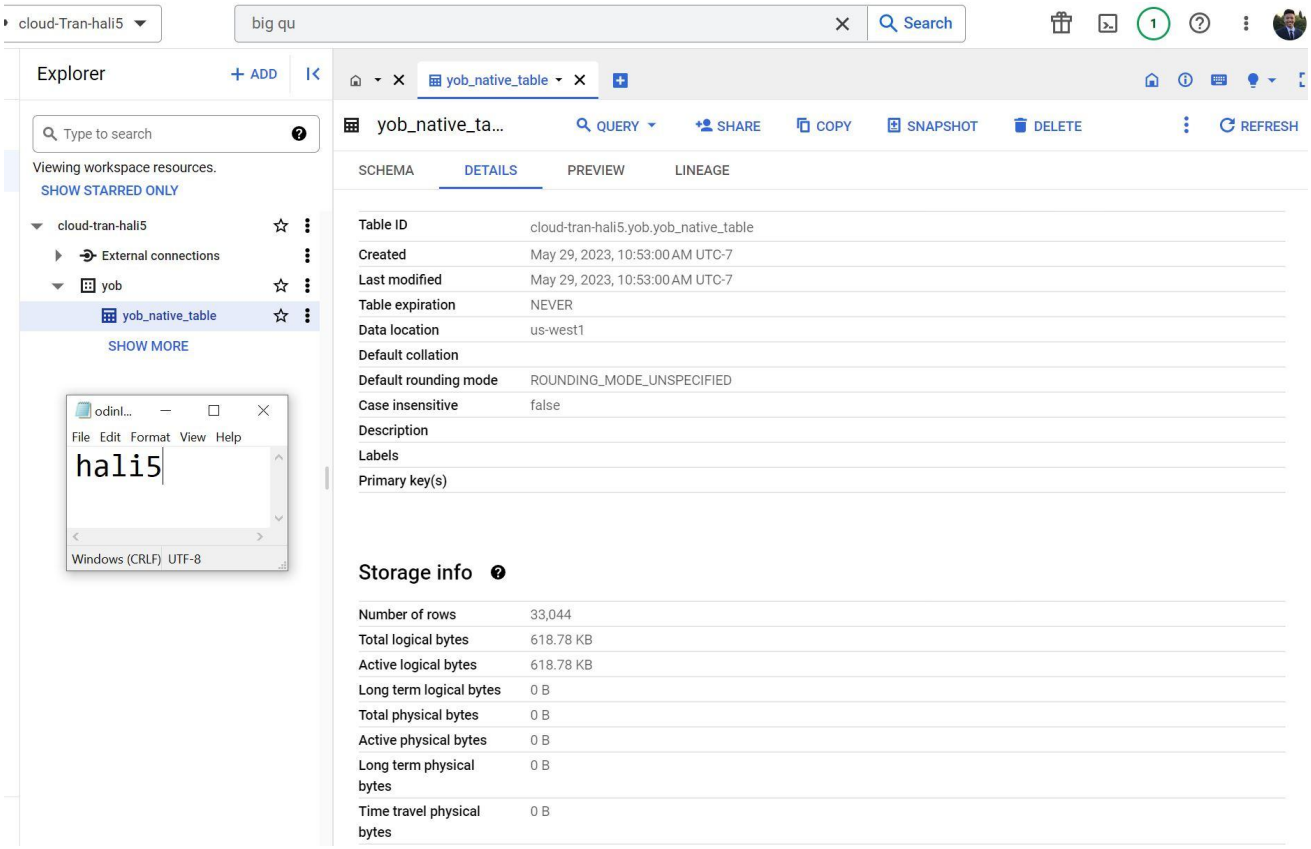
09.1g: BigQuery, BigLake

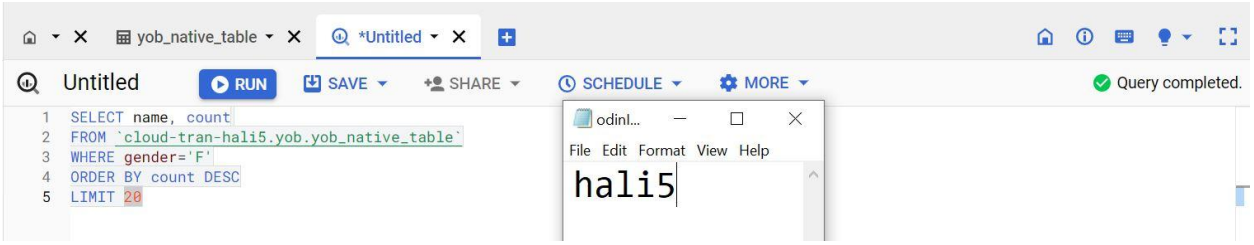
Create dataset

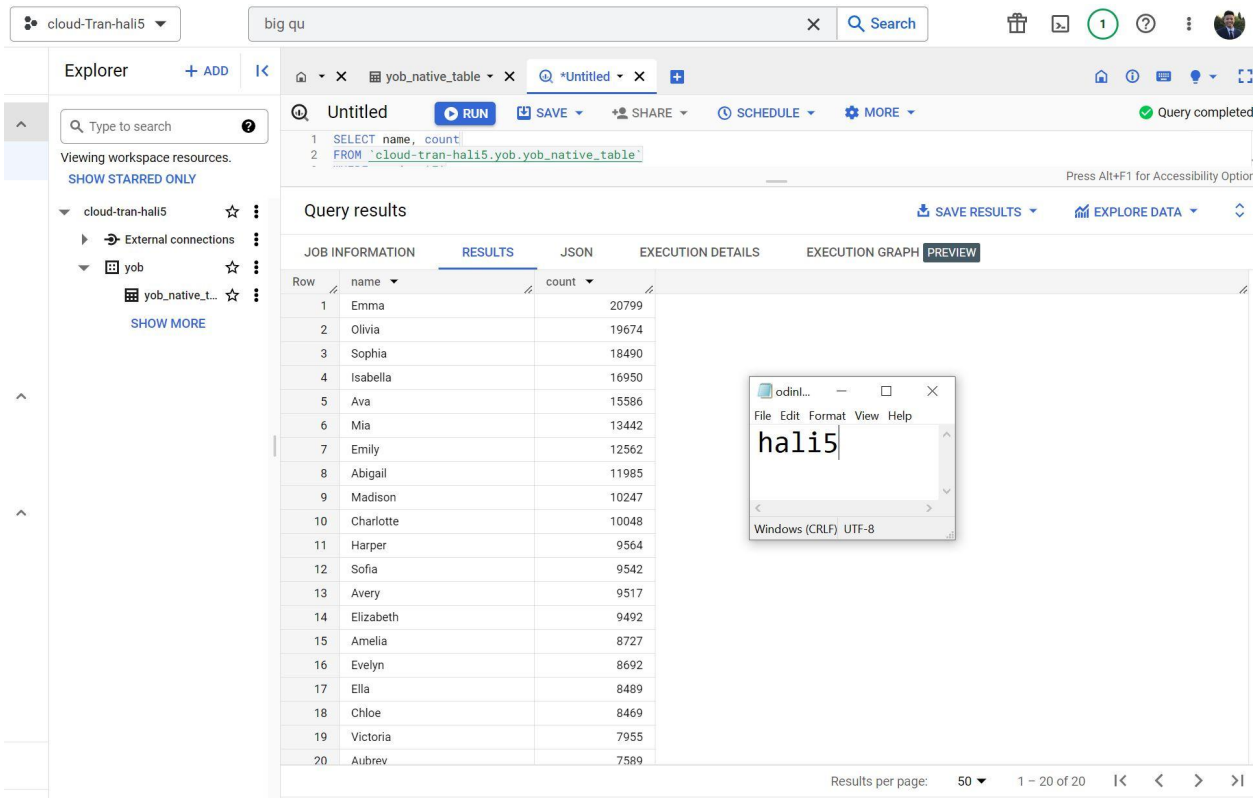
Take a screenshot of the table's details that includes the number of rows in the table.



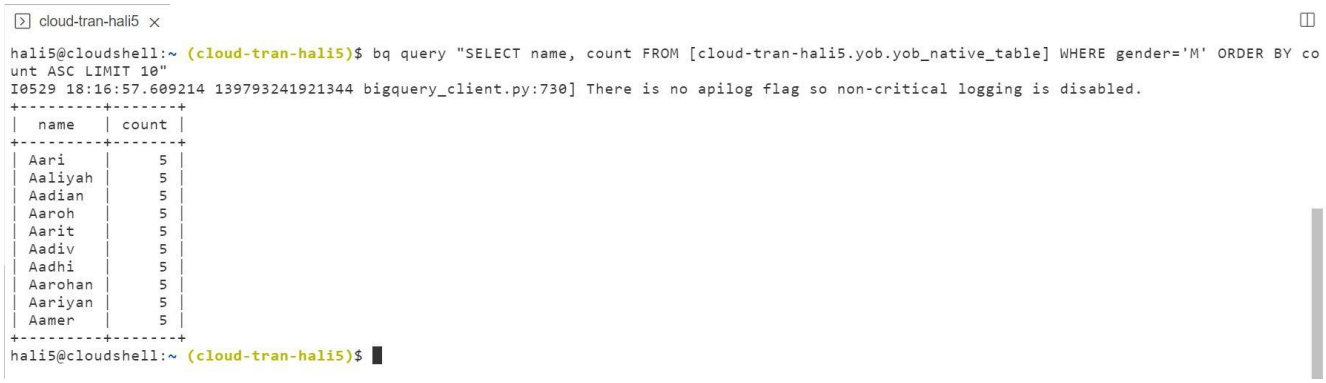
Query data

Screenshot the query results and include it in your lab notebook





Screenshot your results and include it in your lab notebook



Screenshot your results and include it in your lab notebook

```
cloud-tran-hali5 ×

hali5@cloudshell:~ (cloud-tran-hali5)$ bq shell
I0529 18:28:17.552992 140271381657408 bigquery_client.py:730] There is no apilog flag so non-critical logging is disabled.
Welcome to BigQuery! (Type help for more information.)
cloud-tran-hali5> SELECT name, count FROM [cloud-tran-hali5.yob.yob_native_table] WHERE gender='M' ORDER BY count DESC LIMIT 10
+-----------+-------+
|   name    | count |
+-----------+-------+
| Noah      | 19144 |
| Liam      | 18342 |
| Mason     | 17092 |
| Jacob     | 16712 |
| William   | 16687 |
| Ethan     | 15619 |
| Michael   | 15323 |
| Alexander | 15293 |
| James     | 14301 |
| Daniel    | 13829 |
+-----------+-------+
cloud-tran-hali5>
```

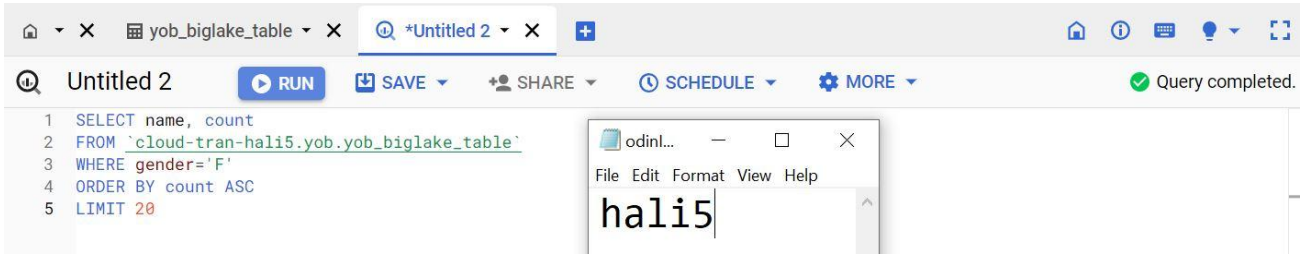
Screenshot your results and include it in your lab notebook

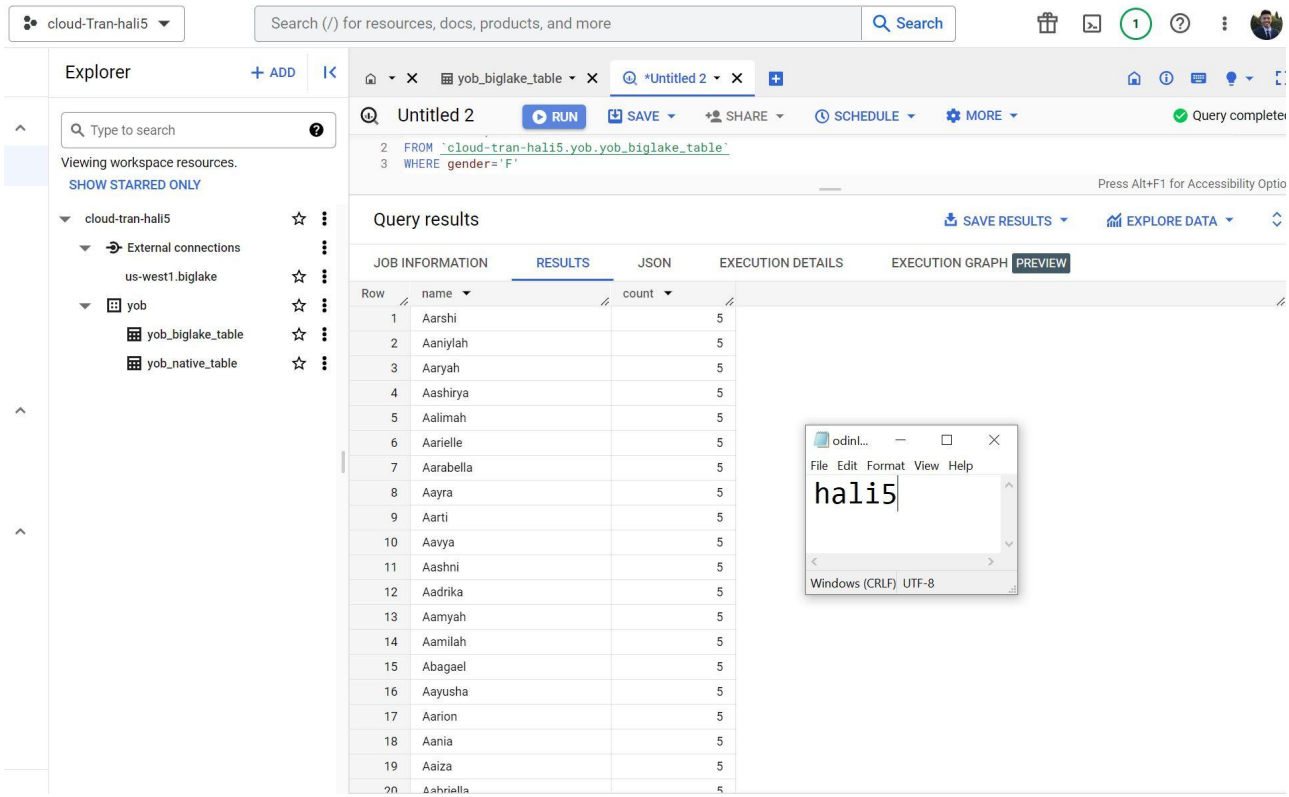
```
cloud-tran-hali5 ×

hali5@cloudshell:~ (cloud-tran-hali5)$ bq shell
I0529 18:34:14.258896 140603490658112 bigquery_client.py:730] There is no apilog flag so non-critical logging is disabled.
Welcome to BigQuery! (Type help for more information.)
cloud-tran-hali5> SELECT name, count FROM [cloud-tran-hali5.yob.yob_native_table] WHERE name='Hali'
+------+-------+
| name | count |
+------+-------+
| Hali |    24 |
+------+-------+
cloud-tran-hali5>
```

Query data

Screenshot the query results and include it in your lab notebook





09.2g: Jupyter Notebooks

BigQuery query

How much less data does this query process compare to the size of the table?

There is roughly 7 times less data in this query process compared to the size of the table. The query will process 3.05 GB compared to the size of the table which is 21.94 GB.

How many twins were born during this time range?

There were 375362 twins born during this time range.

How much lighter on average are they compared to single babies?

On average, twins are roughly 2 pounds lighter than single babies.

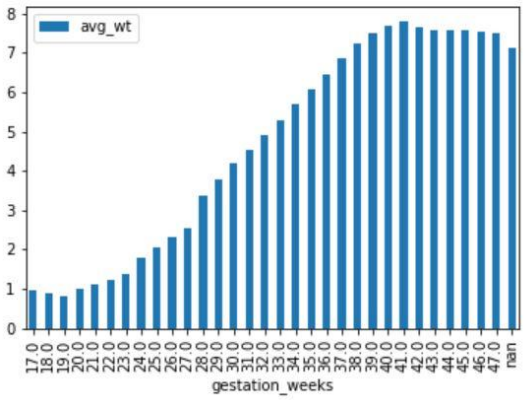
Run queries

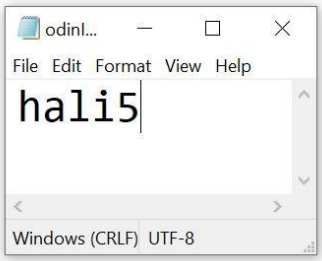
Show the plots generated for the two most important features for your lab notebook.

The two strongest predictors for a newborn baby's weight are plurality and gestation weeks.

```
[8]: df = get_distinct_values('gestation_weeks')
     df.plot(x='gestation_weeks', y='avg_wt', kind='bar')
```

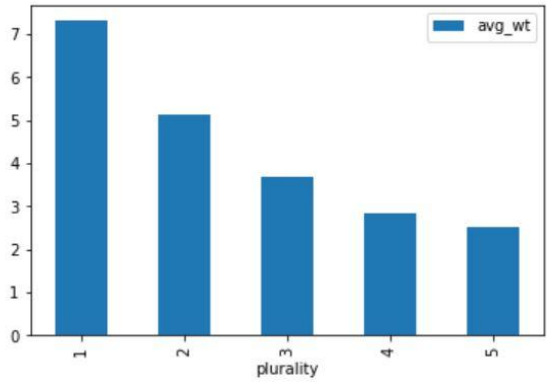
```
[8]: <matplotlib.axes._subplots.AxesSubplot at 0x7fc2d2ed6d90>
```

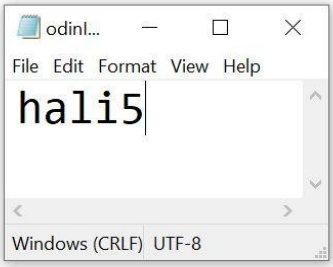





```
[6]: df = get_distinct_values('plurality')
     df.plot(x='plurality', y='avg_wt', kind='bar')
```

```
[6]: <matplotlib.axes._subplots.AxesSubplot at 0x7fc2d806e490>
```





Mobility

What day saw the largest spike in trips to grocery and pharmacy stores?

2020/03/13 saw the largest spike to grocery and pharmacy stores.

On the day the stay-at-home order took effect (3/23/2020), what was the total impact on workplace trips?

The total impact on workplace trips was -49.

Airport traffic

Which three airports were impacted the most in April 2020 (the month when lockdowns became widespread)?

Detroit Metropolitan Wayne County, McCarran International, and San Francisco International were impacted the most in April 2020.

Run the query again using the month of August 2020. Which three airports were impacted the most?

The three airports from the previous question Detroit Metropolitan Wayne County, McCarran International, and San Francisco International were impacted the most in August 2020.

Mortality

What table and columns identify the place name, the starting date, and the number of excess deaths from COVID-19?

The excess_deaths table and its corresponding columns *placename*, *start_date*, and *excess_deaths* identify this.

What table and columns identify the date, county, and deaths from COVID-19?

The us_counties table and its corresponding columns *date*, *county*, and *deaths* identify this.

What table and columns identify the date, state, and confirmed cases of COVID-19?

The us_states table and its corresponding columns *date*, *state_name*, and *deaths* identify this.

What table and columns identify a county code and the percentage of its residents that report they always wear masks?

The mask_use_by_county table and its corresponding columns *country_fips_code* and *always* identify this.

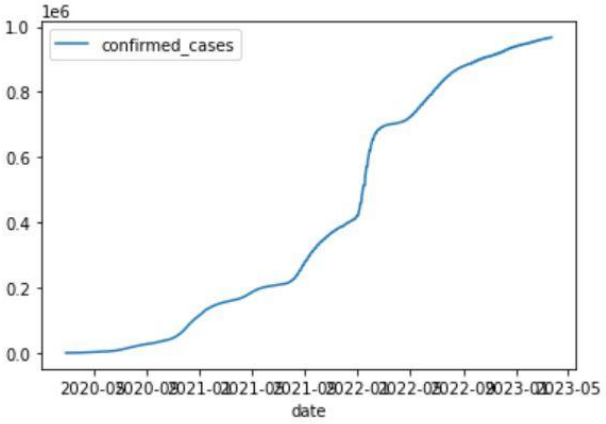
Run example queries

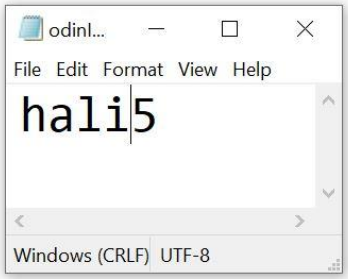
Show a screenshot of the plot and the code used to generate it for your lab notebook

```
[16]: query_string = """
      SELECT date, confirmed_cases
      FROM `bigquery-public-data.covid19_nyt.us_states`
      WHERE state_name = 'Oregon'
      ORDER BY date ASC
      """

      from google.cloud import bigquery
      df = bigquery.Client().query(query_string).to_dataframe()
      df.plot(x='date', y='confirmed_cases', kind='line')

[16]: <matplotlib.axes._subplots.AxesSubplot at 0x7fc2d13b6750>
```





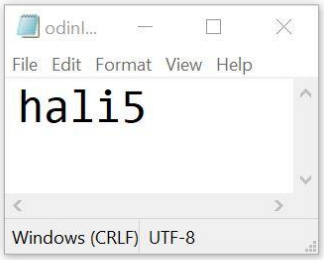
From within your Jupyter notebook, run the query and write code that shows the first 10 states that reached 1000 deaths from COVID-19. Take a screenshot for your lab notebook.

[18]:
```
query_string = """
SELECT state_name, MIN(date) as date_of_1000
FROM `bigquery-public-data.covid19_nyt.us_states`
WHERE deaths > 1000
GROUP BY state_name
ORDER BY date_of_1000 ASC
LIMIT 10
"""

from google.cloud import bigquery
df = bigquery.Client().query(query_string).to_dataframe()
df.head(10)
```

[18]:

	state_name	date_of_1000
0	New York	2020-03-29
1	New Jersey	2020-04-06
2	Michigan	2020-04-09
3	Louisiana	2020-04-14
4	Massachusetts	2020-04-15
5	Illinois	2020-04-16
6	Pennsylvania	2020-04-17
7	Connecticut	2020-04-17
8	California	2020-04-17
9	Florida	2020-04-24



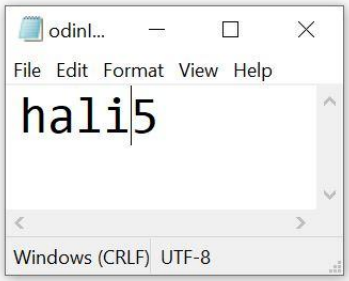
Take a screenshot for your lab notebook of the Top 5 counties and the states they are located in.

[19]:
```
query_string = """
SELECT DISTINCT mu.county_fips_code, mu.always, ct.county, ct.state_name
FROM `bigquery-public-data.covid19_nyt.mask_use_by_county` as mu
LEFT JOIN `bigquery-public-data.covid19_nyt.us_counties` as ct
ON mu.county_fips_code = ct.county_fips_code
ORDER BY mu.always DESC
LIMIT 5
"""

from google.cloud import bigquery
df = bigquery.Client().query(query_string).to_dataframe()
df.head(5)
```

[19]:

	county_fips_code	always	county	state_name
0	06027	0.889	Inyo	California
1	36123	0.884	Yates	New York
2	48229	0.880	Hudspeth	Texas
3	06051	0.880	Mono	California
4	48141	0.877	El Paso	Texas



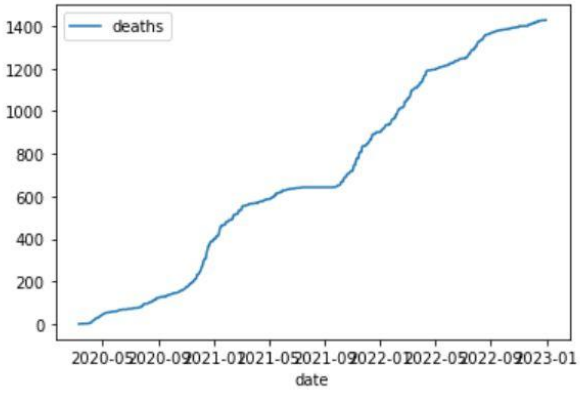
Write queries

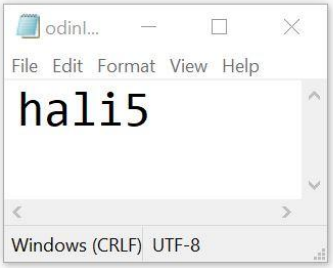
Plot the results and take a screenshot for your lab notebook.

[21]:
```
query_string = """
SELECT date, deaths
FROM `bigquery-public-data.covid19_nyt.us_counties`
WHERE county='Multnomah'
ORDER BY date ASC
"""

from google.cloud import bigquery
df = bigquery.Client().query(query_string).to_dataframe()
df.plot(x='date', y='deaths', kind='line')
```

[21]: `<matplotlib.axes._subplots.AxesSubplot at 0x7fc2d0e3bf90>`



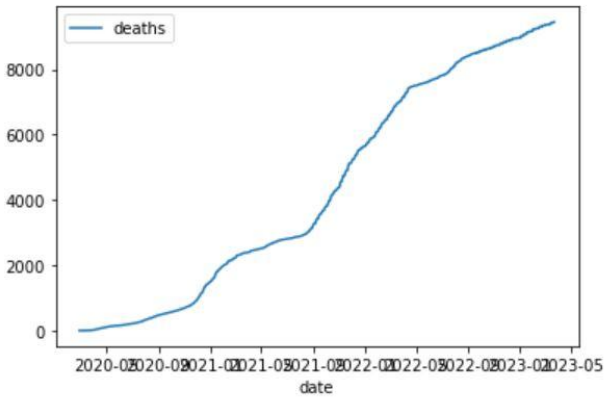


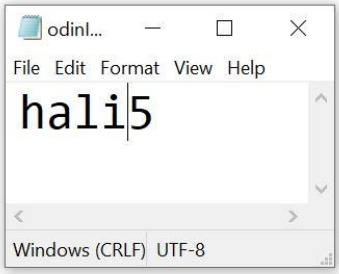
Plot the results and take a screenshot for your lab notebook.

[24]:
```
query_string = """
SELECT date, deaths
FROM `bigquery-public-data.covid19_nyt.us_states`
WHERE state_name='Oregon'
ORDER BY date ASC
"""

from google.cloud import bigquery
df = bigquery.Client().query(query_string).to_dataframe()
df.plot(x='date', y='deaths', kind='line')
```

[24]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x7fc2d05f4e10>
```





09.3g: Dataproc

Run computation

How long did the job take to execute?

The job took 1 minute, 13 seconds to execute.

Examine output.txt and show the estimate of π calculated.

```
hali5@cloudshell:~ (cloud-tran-hali5)$ cat output.txt
Job [ab15ef93313348c69960ba4a2d068c4a] submitted.
Waiting for job output...
23/05/30 05:58:22 INFO org.apache.spark.SparkEnv: Registering MapOutputTracker
23/05/30 05:58:22 INFO org.apache.spark.SparkEnv: Registering BlockManagerMaster
23/05/30 05:58:22 INFO org.apache.spark.SparkEnv: Registering BlockManagerMasterHeartbea
23/05/30 05:58:22 INFO org.apache.spark.SparkEnv: Registering OutputCommitCoordinator
23/05/30 05:58:22 INFO org.sparkproject.jetty.util.log: Logging initialized @3767ms to c
23/05/30 05:58:22 INFO org.sparkproject.jetty.server.Server: jetty-9.4.40.v20210413; bui
23/05/30 05:58:22 INFO org.sparkproject.jetty.server.Server: Started @3873ms
23/05/30 05:58:22 INFO org.sparkproject.jetty.server.AbstractConnector: Started ServerCo
23/05/30 05:58:23 INFO org.apache.hadoop.yarn.client.RMProxy: Connecting to ResourceMana
23/05/30 05:58:23 INFO org.apache.hadoop.yarn.client.AHSProxy: Connecting to Application
23/05/30 05:58:24 INFO org.apache.hadoop.conf.Configuration: resource-types.xml not foun
23/05/30 05:58:24 INFO org.apache.hadoop.yarn.util.resource.ResourceUtils: Unable to fin
23/05/30 05:58:26 INFO org.apache.hadoop.yarn.client.api.impl.YarnClientImpl: Submitted
23/05/30 05:58:27 INFO org.apache.hadoop.yarn.client.RMProxy: Connecting to ResourceMana
23/05/30 05:58:29 INFO com.google.cloud.hadoop.repackaged.gcs.com.google.cloud.hadoop.gc
already exists with desired state.
Pi is roughly 3.1416515514165155
```

The estimate of Pi calculated was 3.1416515514165155.

Run computation again

How long did the job take to execute? How much faster did it take?

The job took 37 seconds to execute. It was nearly 2 times faster than the previous.

Examine output2.txt and show the estimate of π calculated.

```
hali5@cloudshell:~ (cloud-tran-hali5)$ cat output2.txt
Job [656ee2f82afa43cfba775904ba4c847b] submitted.
Waiting for job output...
23/05/30 06:15:45 INFO org.apache.spark.SparkEnv: Registering MapOutputTracker
23/05/30 06:15:45 INFO org.apache.spark.SparkEnv: Registering BlockManagerMaster

23/05/30 06:15:46 INFO org.apache.spark.SparkEnv: Registering BlockManagerMasterHeartbe
23/05/30 06:15:46 INFO org.apache.spark.SparkEnv: Registering OutputCommitCoordinator
23/05/30 06:15:46 INFO org.sparkproject.jetty.util.log: Logging initialized @3232ms to
23/05/30 06:15:46 INFO org.sparkproject.jetty.server.Server: jetty-9.4.40.v20210413; bu
23/05/30 06:15:46 INFO org.sparkproject.jetty.server.Server: Started @3322ms
23/05/30 06:15:46 INFO org.sparkproject.jetty.server.AbstractConnector: Started ServerC
23/05/30 06:15:47 INFO org.apache.hadoop.yarn.client.RMProxy: Connecting to ResourceMan
23/05/30 06:15:47 INFO org.apache.hadoop.yarn.client.AHSProxy: Connecting to Applicatio
23/05/30 06:15:47 INFO org.apache.hadoop.conf.Configuration: resource-types.xml not fou
23/05/30 06:15:47 INFO org.apache.hadoop.yarn.util.resource.ResourceUtils: Unable to fi
23/05/30 06:15:48 INFO org.apache.hadoop.yarn.client.api.impl.YarnClientImpl: Submitted
23/05/30 06:15:49 INFO org.apache.hadoop.yarn.client.RMProxy: Connecting to ResourceMan
23/05/30 06:15:51 INFO com.google.cloud.hadoop.repackaged.gcs.com.google.cloud.hadoop.g
already exists with desired state.
Pi is roughly 3.1416726314167263
```

The estimate of Pi calculated was 3.1416726314167263.

09.4g: Dataflow

Beam code

Where is the input taken from by default?

The input by default is taken from `../javahelp/src/main/java/com/google/cloud/training/dataanalyst/javahelp/`

Where does the output go by default?

The output by default is taken from `/tmp/output`

Examine both the getPackages() function and the splitPackageName() function. What operation does the 'PackageUse()' transform implement?

The PackageUse transform takes a line of a code and a keyword and applies the getPackages function to extract the package name. It then yields (package, 1) pairs for each package used in the line of code.

Look up Beam's CombinePerKey. What operation does the TotalUse operation implement?

The TotalUse implements the operation to calculate the total count of each package used in the Java source code by taking the output from the PackageUse transformation. It utilizes CombinePerKey(sum) to calculate the total count of packages.

Which operations correspond to a "Map"?

GetJava, GetImports, and PackageUse operations correspond to a Map.

Which operation corresponds to a "Shuffle-Reduce"?

CombinePerKey operation corresponds to Shuffle-Reduce.

Which operation corresponds to a "Reduce"?

Top_5 operation corresponds to Reduce.

Run pipeline locally

Take a screenshot of its contents

```
(env) hali5@cloudshell:/tmp (cloud-tran-hali5)$ cat output-00000-of-00001
[('org', 45), ('org.apache', 44), ('org.apache.beam', 44), ('org.apache.beam.sdk', 43), ('org.apache.beam.sdk.transforms', 16)]
(env) hali5@cloudshell:/tmp (cloud-tran-hali5)$
```

Explain what the data in this output file corresponds to based on your understanding of the program.

The data in the output file corresponds to the top 5 Java packages that are used the most frequently in the Java source code that is inputted into the pipeline.

Dataflow Lab #2 (Word count)

What are the names of the stages in the pipeline?

The names of the states in the pipeline are Read, Split, PairWithOne, GroupAndSum, and Format.

Describe what each stage does.

Read: Reads the input text file
Split: Applies WordExtractingDoFn function transformation to split each line of the input into individual words.
PairWithOne: Uses Map transform to convert each individual word to a key value pair.
GroupAndSum: Uses the CombinePerKey transformation to group the key-value pairs by key and sum the corresponding values.
Format: Uses the MapTuple function to format the counts into a PCollection of strings. It formats the key-value pairs.
Write: Writes the PCollection to an output file.

Run code locally

Use wc with an appropriate flag to determine the number of unique words in King Lear.

```
(env) hali5@cloudshell:~/training-data-analyst/courses/machine_learning/deepdive/04_features/dataflow/python (cloud-tran-hali5)$ wc -l outputs-00000-of-0
0001
4784 outputs-00000-of-00001
(env) hali5@cloudshell:~/training-data-analyst/courses/machine_learning/deepdive/04_features/dataflow/python (cloud-tran-hali5)$
```

Use sort with appropriate flags to perform a *numeric* sort on the *key field* containing the count for each word in *descending* order. Pipe the output into head to show the top 3 words in King Lear and the number of times they appear.

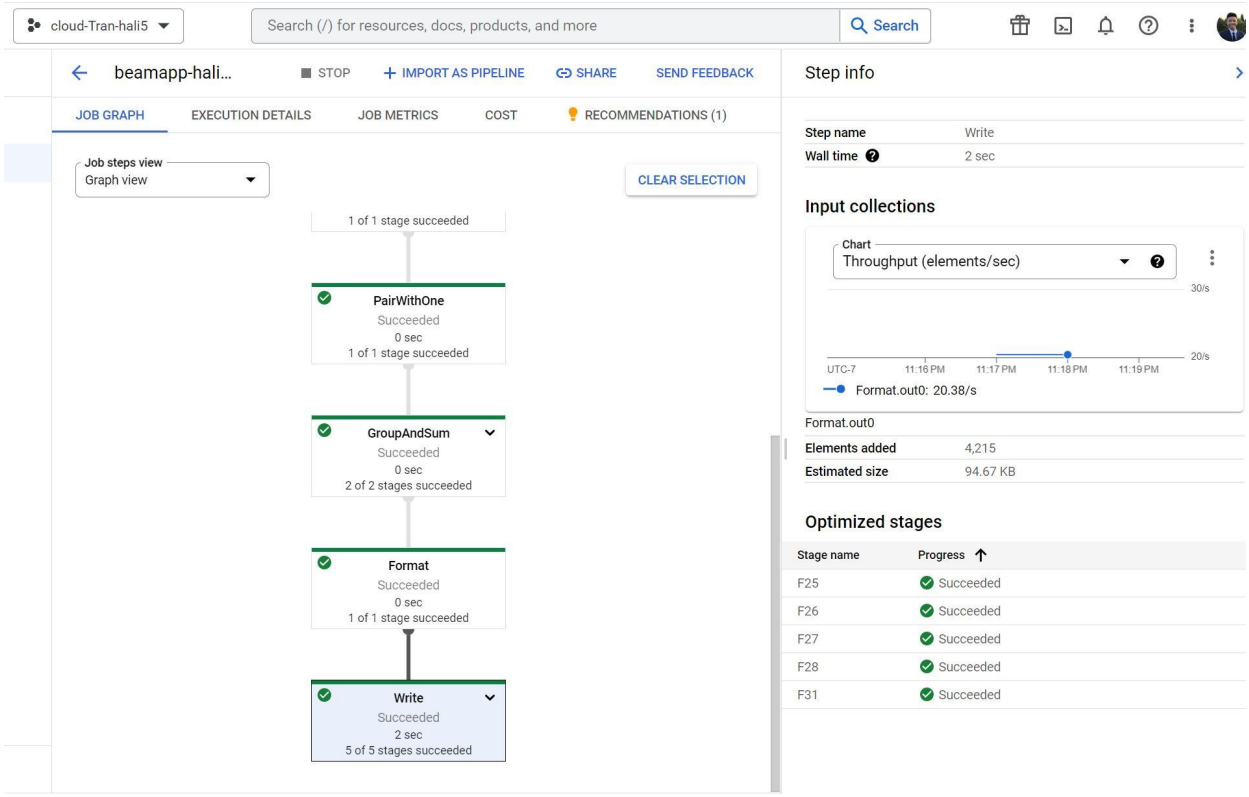
```
(env) hali5@cloudshell:~/training-data-analyst/courses/machine_learning/deepdive/04_features/dataflow/python (cloud-tran-hali5)$ sort -k2 -nr outputs-00000-of-00001 | head -n 3
the: 786
I: 622
and: 594
(env) hali5@cloudshell:~/training-data-analyst/courses/machine_learning/deepdive/04_features/dataflow/python (cloud-tran-hali5)$
```

Use the previous method to show the top 3 words in King Lear, case-insensitive, and the number of times they appear.

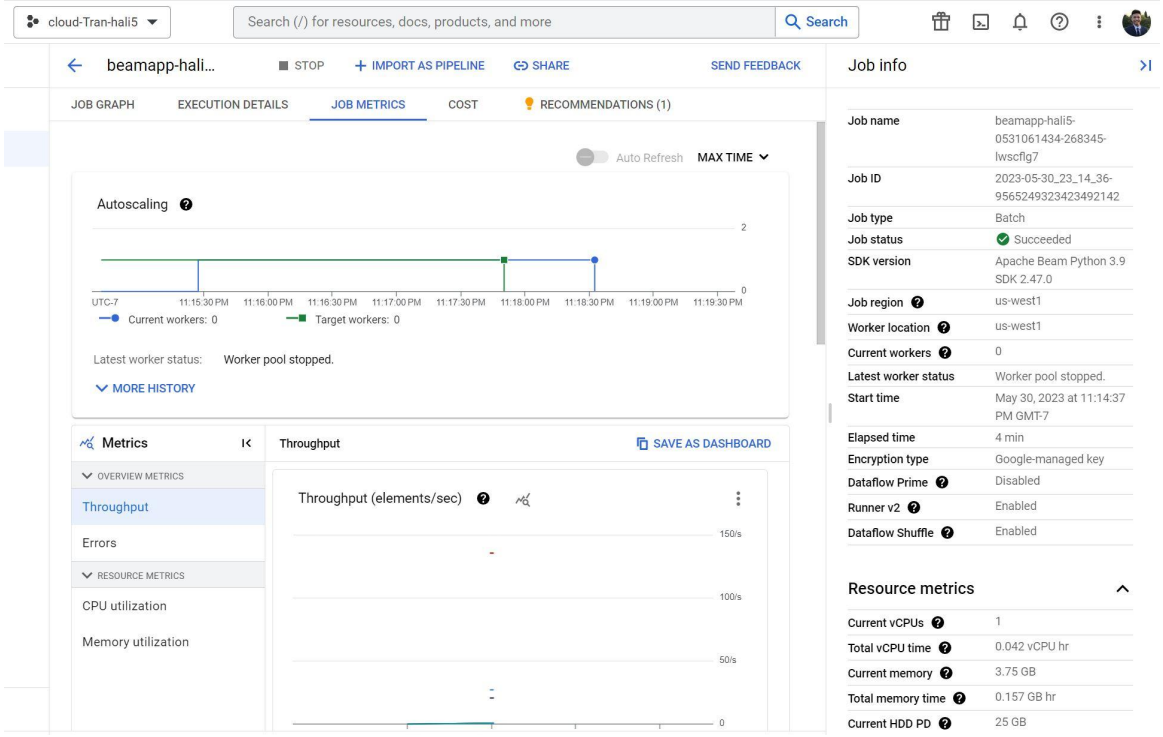
```
(env) hali5@cloudshell:~/training-data-analyst/courses/machine_learning/deepdive/04_features/dataflow/python (cloud-tran-hali5)$ sort -k2 -nr outputs-00000-of-00001 | head -n 3
the: 908
and: 738
i: 622
(env) hali5@cloudshell:~/training-data-analyst/courses/machine_learning/deepdive/04_features/dataflow/python (cloud-tran-hali5)$
```

Run code using Dataflow runner

The part of the job graph that has taken the longest time to complete.



The auto scaling graph shows when the worker was created and stopped.



Examine the output directory in Cloud Storage. How many files has the final write stage in the pipeline created?

The final write stage in the pipeline created a single file named outputs.

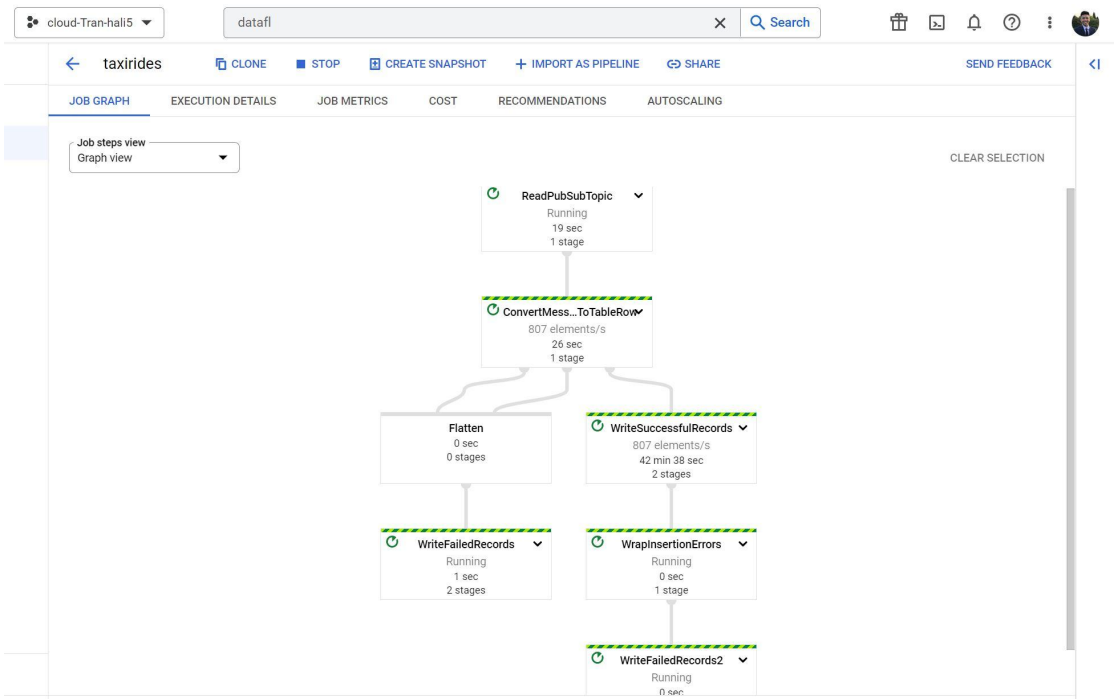
View raw data from PubSub

Take a screenshot listing the different fields of this object.

```
hali5@cloudshell:~ (cloud-tran-hali5)$ gcloud pubsub subscriptions pull taxisub --auto-ack
DATA: {"ride_id":"7f9aba63-f93d-4cf9-89eb-23770b940bd9","point_idx":683,"latitude":40.736850000000004,
"longitude":-73.93122000000001,"timestamp":"2023-05-31T02:33:07.46506-04:00","meter_reading":12.459263
,"meter_increment":0.018241966,"ride_status":"enroute","passenger_count":1}
MESSAGE_ID: 7861017045290965
ORDERING_KEY:
ATTRIBUTES: ts=2023-05-31T02:33:07.46506-04:00
DELIVERY_ATTEMPT:
ACK_STATUS: SUCCESS
hali5@cloudshell:~ (cloud-tran-hali5)$
```

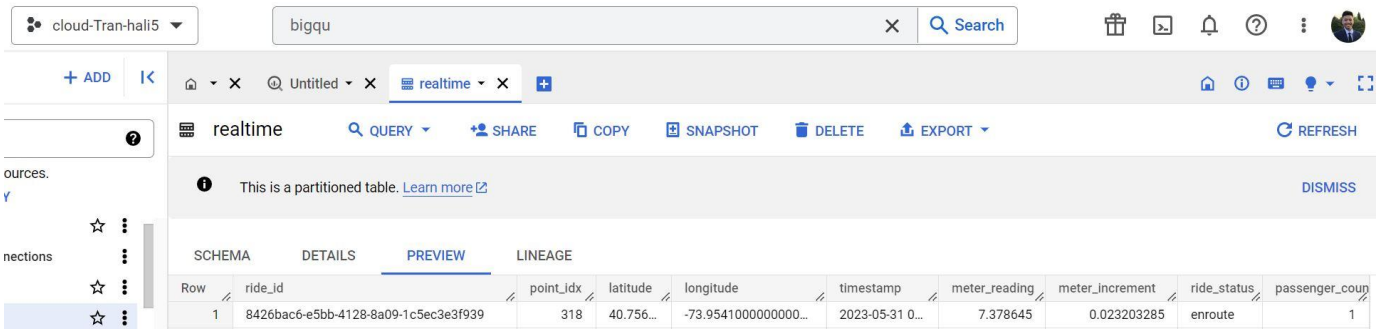
Run Dataflow job from template

Take a screenshot of the pipeline that includes its stages and the number of elements per second being handled by individual stages.

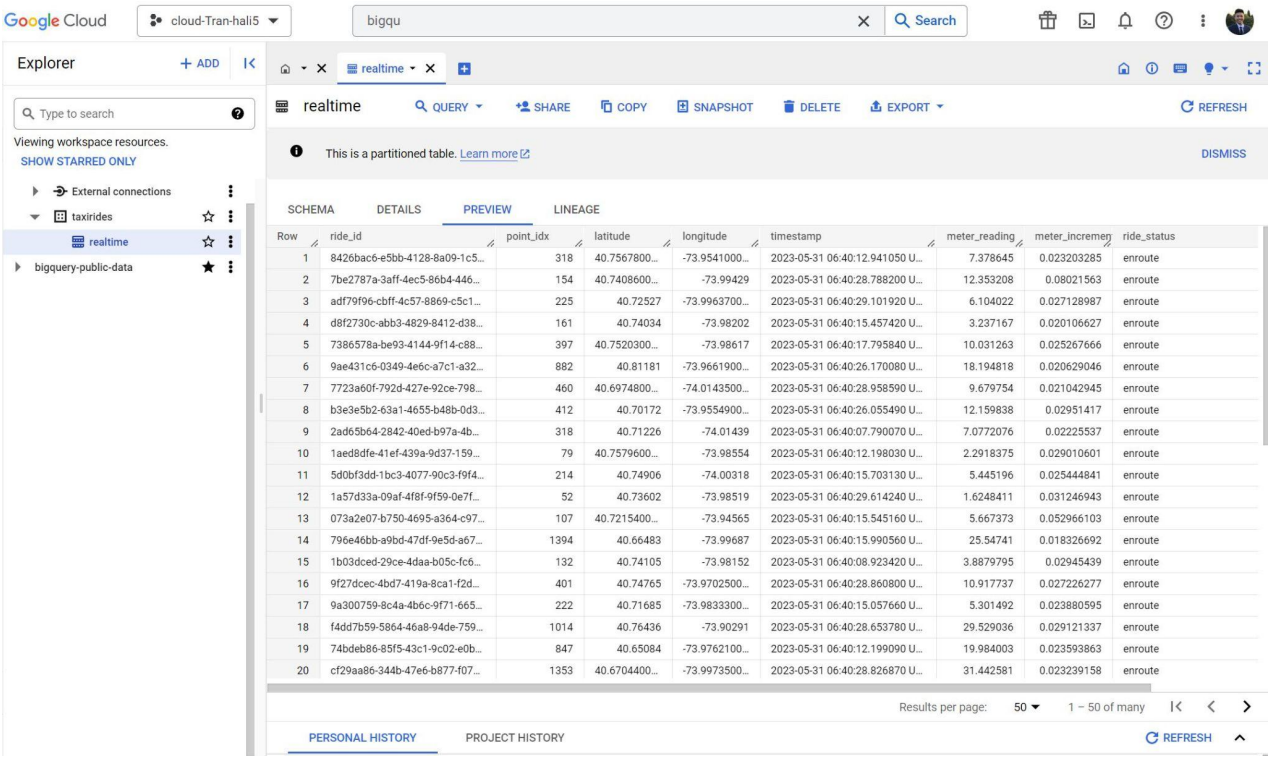


Query data in BigQuery

Take a screenshot showing the number of passengers and the amount paid for the first ride



Take a screenshot showing the estimated number of rows in the table.



Take a screenshot showing the per-minute number of rides, passengers, and revenue for the data collected

realtime | *Untitled

Untitled | RUN | SAVE | SHARE | SCHEDULE | MORE | Query completed.

```
1  SELECT
2    FORMAT_TIMESTAMP("%R", timestamp, "America/Los_Angeles") AS minute,
3    COUNT(DISTINCT ride_id) AS total_rides,
4    SUM(passenger_count) AS total_passengers,
5    SUM(meter_reading) AS total_revenue
6  FROM
7    taxirides.realtime
8  WHERE
9    ride_status = 'dropoff'
10 GROUP BY
11   minute
12 ORDER BY
13   minute ASC
```

Press Alt+F1 for Accessibility Options

Query results | SAVE RESULTS | EXPLORE DATA

JOB INFORMATION | RESULTS | JSON | EXECUTION DETAILS | EXECUTION GRAPH | PREVIEW

Row	minute	total_rides	total_passengers	total_revenue
1	23:40	82	142	1321.6899944000004
2	23:41	65	93	1010.3699938
3	23:42	89	146	1524.6500007000004
4	23:43	102	186	1777.8699993999999
5	23:44	68	97	1018.9399953000001
6	23:45	28	55	352.99999970000005

Data visualization

Take a screenshot showing the plot for your data for your lab notebook