09.2g: BigQuery, JupyterLab **2**
3. Create dataset 3
Take a screenshot of the table's details that includes the number of rows in the table. 3
4. Query data 4
Screenshot your results and include it in your lab notebook 4
6. BigQuery query 5
How many twins were born during this time? 5
9. Run queries 5
Show the plots generated for the two most important features for your lab notebook 6
10. BigQuery, Notebooks Lab #3 (COVID-19 Mobility) 6
What day saw the largest spike in trips to grocery and pharmacy stores? 7
On the day the stay-at-home order took effect (3/23/2020), what was the total impact on workplace trips? 7
11. BigQuery, Notebooks Lab #4 (COVID-19 NYT) 7
What table and columns identify the place name, the starting date, and the number of excess deaths from COVID-19? 8
What table and columns identify the date, county, and deaths from COVID-19? 8
What table and columns identify the date, state, and confirmed cases of COVID-19? 8
What table and columns identify a county code and the percentage of its residents that report they always wear masks? 8
12. Run example queries 8
Show a screenshot of the plot and the code used to generate it for your lab notebook 8
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. 8
Take a screenshot for your lab notebook of the Top 5 counties and the states they are located in. 8
13. Write queries 8
Plot the results and take a screenshot for your lab notebook. 8
Plot the results and take a screenshot for your lab notebook. 8

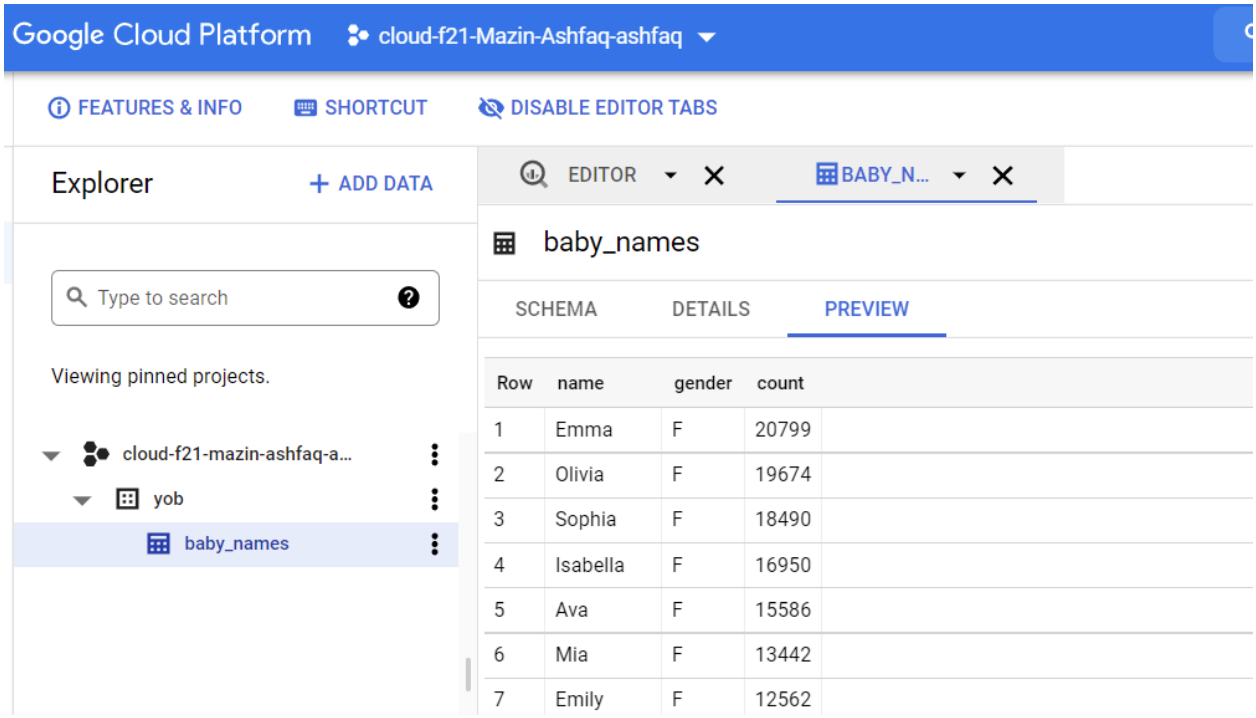
1. Dataproc Lab #1 (π) **8**
6. Run computation 8
How long did the job take to execute? 8
Examine output.txt and show the estimate of π calculated. 8
8. Run computation again 8
For your lab notebook: 8
How long did the job take to execute? How much faster did it take? 8
Examine output2.txt and show the estimate of π calculated. 9
12. Beam code 10
Where is the input taken from by default? 10
Where does the output go by default? 10

Examine both the getPackages() function and the splitPackageName() function. What operation does the 'PackageUse()' transform implement? 10
Look up Beam's CombinePerKey. What operation does the TotalUse operation implement? 10
Which operations correspond to a "Map"? 10
Which operation corresponds to a "Shuffle-Reduce"? 10
Which operation corresponds to a "Reduce"? 10
13. Run pipeline locally 10
Take a screenshot of its contents 10
Explain what the data in this output file corresponds to based on your understanding of the program. 10
14. Dataflow Lab #2 (Word count) 10
What are the names of the stages in the pipeline? 10
Describe what each stage does. 10
15. Run code locally 10
Use wc with an appropriate flag to determine the number of unique words in King Lear. 10
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 10
Use the previous method to show the top 3 words in King Lear, case-insensitive, and the number of times they appear. 10
18. Run code using Dataflow runner 10
The part of the job graph that has taken the longest time to complete. 10
The autoscaling graph showing when the worker was created and stopped. 10
Examine the output directory in Cloud Storage. How many files has the final write stage in the pipeline created? 10

09.2g: BigQuery, JupyterLab

3. Create dataset

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



4. Query data

- Screenshot your results and include it in your lab notebook

```
1   SELECT name, count
2   FROM `cloud-f21-mazin-ashfaq-ashfaq.yob.baby_names`
3   WHERE gender='F'
```

Query results SAVE RESULTS EXPLORE DATA

Query complete (0.2 sec elapsed, 618.8 KB processed)

Job information | Results | JSON | Execution details

Row	name	count
1	Emma	20799
2	Olivia	19674
3	Sophia	18490
4	Isabella	16950
5	Ava	15586
6	Mia	13442
7	Emily	12562
8	Abigail	11985
9	Madison	10247
10	Charlotte	10048

- Screenshot your results and include it in your lab notebook

```
Waiting on bqjob_r59c6f045a163cb08_0000017d6f81feca_1 ... (0s) Current status: DONE
+---------+-------+
|  name   | count |
+---------+-------+
| Aari    |     5 |
| Aaliyah |     5 |
| Aadian  |     5 |
| Aaroh   |     5 |
| Aarit   |     5 |
| Aadiv   |     5 |
| Aadhi   |     5 |
| Aarohan |     5 |
| Aariyan |     5 |
| Aamer   |     5 |
+---------+-------+
ashfaq@cloudshell:~ (cloud-f21-mazin-ashfaq-ashfaq)$
```

- Screenshot your results and include it in your lab notebook

```
cloud-f21-mazin-ashfaq-ashfaq> SELECT name, count FROM [cloud-f21-mazin-ashfaq-ashfaq.yob.baby_names] WHERE gender='M' ORDER BY count DESC LIMIT 10
Waiting on bqjob_r67946b882f07446d_0000017d6f83f355_1 ... (0s) Current status: DONE
+-----------+-------+
|   name    | count |
+-----------+-------+
| Noah      | 19144 |
| Liam      | 18342 |
| Mason     | 17092 |
| Jacob     | 16712 |
| William   | 16687 |
| Ethan     | 15619 |
| Michael   | 15323 |
| Alexander | 15293 |
| James     | 14301 |
| Daniel    | 13829 |
+-----------+-------+
cloud-f21-mazin-ashfaq-ashfaq>
```

- Screenshot your results and include it in your lab notebook

```
Waiting on bqjob_r4c773a0b04a4ec88_0000017d6f85f62a_1 ... (0s) Current status: DONE
+--------+-------+
| name   | count |
+--------+-------+
| Mazin  |    32 |
+--------+-------+
cloud-f21-mazin-ashfaq-ashfaq>
```

6. BigQuery query

- How many twins were born during this time?

  11,779,637

Query complete (0.6 sec elapsed, 3.1 GB processed)

Job information | Results | JSON | Execution details

Row	plurality	num_babies	avg_wt
1	1	11757058	7.345780731912613
2	4	1407	2.851986586921067
3	3	20933	3.7097715771323494
4	5	239	2.6967018620810164

PERSONAL HISTORY | PROJECT HISTORY | SAVED QUERIES

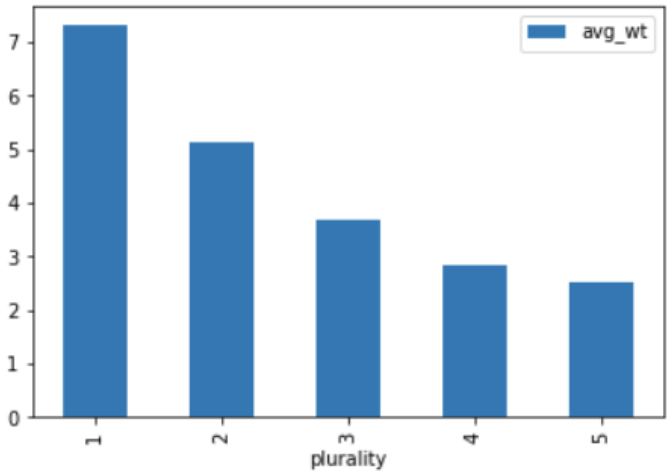
```
arted.
 set to cloud-f21-mazin-ashfaq-ashfaq.
```

9. Run queries

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

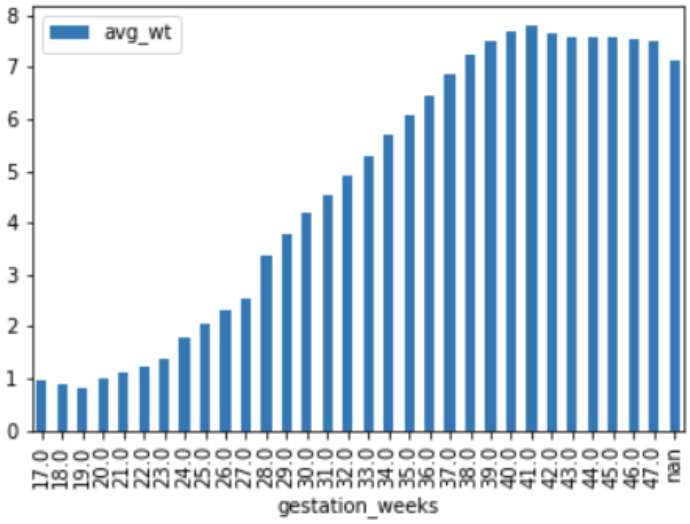
[7]:
```
df = get_distinct_values('plurality')
df.plot(x='plurality', y='avg_wt', kind='bar')
```

[7]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x7f8f00045a10>
```



[9]:
```
df = get_distinct_values('gestation_weeks')
df.plot(x='gestation_weeks', y='avg_wt', kind='bar')
```

[9]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x7f8efb716f90>
```



10. BigQuery, Notebooks Lab #3 (COVID-19 Mobility)

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

  3/12

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

  -49%

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

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

11. BigQuery, Notebooks Lab #4 (COVID-19 NYT)

- What table and columns identify the place name, the starting date, and the number of excess deaths from COVID-19?
- What table and columns identify the date, county, and deaths from COVID-19?
- What table and columns identify the date, state, and confirmed cases of COVID-19?
- What table and columns identify a county code and the percentage of its residents that report they always wear masks?

12. Run example queries

- Show a screenshot of the plot and the code used to generate it for your lab notebook
- 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.
- Take a screenshot for your lab notebook of the Top 5 counties and the states they are located in.

13. Write queries

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

1. Dataproc Lab #1 (π)

6. Run computation

- How long did the job take to execute?

  Less than a second

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

  Pi is roughly 3.1414761514147616

8. Run computation again

For your lab notebook:

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

  Looks like I messed up the first run timer, but this time it took about 51 seconds.

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

  Pi is roughly 3.141441591414416

12. Beam code

- Where is the input taken from by default?
- Where does the output go by default?
- Examine both the `getPackages()` function and the `splitPackageName()` function. What operation does the '`PackageUse()`' transform implement?
- Look up Beam's `CombinePerKey`. What operation does the `TotalUse` operation implement?
- Which operations correspond to a "Map"?
- Which operation corresponds to a "Shuffle-Reduce"?
- Which operation corresponds to a "Reduce"?

13. Run pipeline locally

- Take a screenshot of its contents
- Explain what the data in this output file corresponds to based on your understanding of the program.

14. Dataflow Lab #2 (Word count)

- What are the names of the stages in the pipeline?
- Describe what each stage does.

15. Run code locally

- Use `wc` with an appropriate flag to determine the number of unique words in King Lear.
- 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
- Use the previous method to show the top 3 words in King Lear, case-insensitive, and the number of times they appear.

18. Run code using Dataflow runner

- The part of the job graph that has taken the longest time to complete.
- The autoscaling graph showing 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?