Exploring and visualizing the baby weight dataset

- By - Gautam Sharma

Create ML dataset by sampling using BigQuery

Let's sample the BigQuery data to create smaller datasets.

In [1]:
```
# Create SQL query using natality data after the year 2000
query = """
SELECT
    weight_pounds,
    is_male,
    mother_age,
    plurality,
    gestation_weeks,
    ABS(FARM_FINGERPRINT(CONCAT(CAST(YEAR AS STRING), CAST(month AS STRING))))
FROM
    publicdata.samples.natality
WHERE year > 2000
"""
```

In [2]:
```
pip install datalab
```

```
Requirement already satisfied: datalab in /opt/conda/lib/python3.7/site-packages
(1.2.0)
Requirement already satisfied: pyyaml>=3.11 in /opt/conda/lib/python3.7/site-pac
kages (from datalab) (5.4.1)
Requirement already satisfied: requests>=2.9.1 in /opt/conda/lib/python3.7/site-
packages (from datalab) (2.25.1)
Requirement already satisfied: plotly>=1.12.5 in /opt/conda/lib/python3.7/site-p
ackages (from datalab) (5.2.2)
Requirement already satisfied: google-cloud-monitoring==0.31.1 in /opt/conda/li
b/python3.7/site-packages (from datalab) (0.31.1)
Requirement already satisfied: configparser>=3.5.0 in /opt/conda/lib/python3.7/s
ite-packages (from datalab) (5.0.2)
Requirement already satisfied: ipykernel>=4.5.2 in /opt/conda/lib/python3.7/site
-packages (from datalab) (6.2.0)
Requirement already satisfied: oauth2client>=2.2.0 in /opt/conda/lib/python3.7/s
ite-packages (from datalab) (4.1.3)
Requirement already satisfied: seaborn>=0.7.0 in /opt/conda/lib/python3.7/site-p
ackages (from datalab) (0.11.2)
Requirement already satisfied: python-dateutil>=2.5.0 in /opt/conda/lib/python3.
7/site-packages (from datalab) (2.8.2)
Requirement already satisfied: pandas>=0.22.0 in /opt/conda/lib/python3.7/site-p
ackages (from datalab) (1.3.2)
Requirement already satisfied: pandas-profiling==1.4.0 in /opt/conda/lib/python
3.7/site-packages (from datalab) (1.4.0)
Requirement already satisfied: pytz>=2015.4 in /opt/conda/lib/python3.7/site-pac
kages (from datalab) (2021.1)
Requirement already satisfied: six>=1.10.0 in /opt/conda/lib/python3.7/site-pack
ages (from datalab) (1.16.0)
```

```
Requirement already satisfied: scikit-learn>=0.18.2 in /opt/conda/lib/python3.7/
site-packages (from datalab) (0.24.2)
Requirement already satisfied: httplib2>=0.10.3 in /opt/conda/lib/python3.7/site
-packages (from datalab) (0.19.1)
Requirement already satisfied: future>=0.16.0 in /opt/conda/lib/python3.7/site-p
ackages (from datalab) (0.18.2)
Requirement already satisfied: google-auth-httplib2>=0.0.2 in /opt/conda/lib/pyt
hon3.7/site-packages (from datalab) (0.1.0)
Requirement already satisfied: google-api-core>=1.10.0 in /opt/conda/lib/python
3.7/site-packages (from datalab) (1.31.1)
Requirement already satisfied: jsonschema>=2.6.0 in /opt/conda/lib/python3.7/sit
e-packages (from datalab) (3.2.0)
Requirement already satisfied: scikit-image>=0.13.0 in /opt/conda/lib/python3.7/
site-packages (from datalab) (0.18.2)
Requirement already satisfied: urllib3>=1.22 in /opt/conda/lib/python3.7/site-pa
ckages (from datalab) (1.26.6)
Requirement already satisfied: mock>=2.0.0 in /opt/conda/lib/python3.7/site-pack
ages (from datalab) (2.0.0)
Requirement already satisfied: google-api-python-client>=1.6.2 in /opt/conda/li
b/python3.7/site-packages (from datalab) (2.17.0)
Requirement already satisfied: psutil>=4.3.0 in /opt/conda/lib/python3.7/site-pa
ckages (from datalab) (5.8.0)
Requirement already satisfied: matplotlib>=1.4 in /opt/conda/lib/python3.7/site-
packages (from pandas-profiling==1.4.0->datalab) (3.4.3)
Requirement already satisfied: jinja2>=2.8 in /opt/conda/lib/python3.7/site-pack
ages (from pandas-profiling==1.4.0->datalab) (2.11.3)
Requirement already satisfied: google-auth<2.0dev,>=1.25.0 in /opt/conda/lib/pyt
hon3.7/site-packages (from google-api-core>=1.10.0->datalab) (1.34.0)
Requirement already satisfied: setuptools>=40.3.0 in /opt/conda/lib/python3.7/si
te-packages (from google-api-core>=1.10.0->datalab) (57.4.0)
Requirement already satisfied: packaging>=14.3 in /opt/conda/lib/python3.7/site-
packages (from google-api-core>=1.10.0->datalab) (21.0)
Requirement already satisfied: googleapis-common-protos<2.0dev,>=1.6.0 in /opt/c
onda/lib/python3.7/site-packages (from google-api-core>=1.10.0->datalab) (1.53.
0)
Requirement already satisfied: protobuf>=3.12.0 in /opt/conda/lib/python3.7/site
-packages (from google-api-core>=1.10.0->datalab) (3.16.0)
Requirement already satisfied: grpcio<2.0dev,>=1.29.0 in /opt/conda/lib/python3.
7/site-packages (from google-api-core>=1.10.0->datalab) (1.38.1)
Requirement already satisfied: uritemplate<4dev,>=3.0.0 in /opt/conda/lib/python
3.7/site-packages (from google-api-python-client>=1.6.2->datalab) (3.0.1)
Requirement already satisfied: rsa<5,>=3.1.4 in /opt/conda/lib/python3.7/site-pa
ckages (from google-auth<2.0dev,>=1.25.0->google-api-core>=1.10.0->datalab) (4.
7.2)
Requirement already satisfied: cachetools<5.0,>=2.0.0 in /opt/conda/lib/python3.
7/site-packages (from google-auth<2.0dev,>=1.25.0->google-api-core>=1.10.0->data
lab) (4.2.2)
Requirement already satisfied: pyasn1-modules>=0.2.1 in /opt/conda/lib/python3.
7/site-packages (from google-auth<2.0dev,>=1.25.0->google-api-core>=1.10.0->data
lab) (0.2.7)
Requirement already satisfied: pyparsing<3,>=2.4.2 in /opt/conda/lib/python3.7/s
ite-packages (from httplib2>=0.10.3->datalab) (2.4.7)
Requirement already satisfied: matplotlib-inline<0.2.0,>=0.1.0 in /opt/conda/li
b/python3.7/site-packages (from ipykernel>=4.5.2->datalab) (0.1.2)
Requirement already satisfied: argcomplete>=1.12.3 in /opt/conda/lib/python3.7/s
ite-packages (from ipykernel>=4.5.2->datalab) (1.12.3)
Requirement already satisfied: debugpy<2.0,>=1.0.0 in /opt/conda/lib/python3.7/s
ite-packages (from ipykernel>=4.5.2->datalab) (1.4.1)
Requirement already satisfied: ipython<8.0,>=7.23.1 in /opt/conda/lib/python3.7/
site-packages (from ipykernel>=4.5.2->datalab) (7.26.0)
Requirement already satisfied: importlib-metadata<5 in /opt/conda/lib/python3.7/
```

```
site-packages (from ipykernel>=4.5.2->datalab) (4.6.4)
Requirement already satisfied: jupyter-client<8.0 in /opt/conda/lib/python3.7/si
te-packages (from ipykernel>=4.5.2->datalab) (6.1.12)
Requirement already satisfied: tornado<7.0,>=4.2 in /opt/conda/lib/python3.7/sit
e-packages (from ipykernel>=4.5.2->datalab) (6.1)
Requirement already satisfied: traitlets<6.0,>=4.1.0 in /opt/conda/lib/python3.
7/site-packages (from ipykernel>=4.5.2->datalab) (5.0.5)
Requirement already satisfied: zipp>=0.5 in /opt/conda/lib/python3.7/site-packag
es (from importlib-metadata<5->ipykernel>=4.5.2->datalab) (3.5.0)
Requirement already satisfied: typing-extensions>=3.6.4 in /opt/conda/lib/python
3.7/site-packages (from importlib-metadata<5->ipykernel>=4.5.2->datalab) (3.10.
0.0)
Requirement already satisfied: jedi>=0.16 in /opt/conda/lib/python3.7/site-packa
ges (from ipython<8.0,>=7.23.1->ipykernel>=4.5.2->datalab) (0.18.0)
Requirement already satisfied: pygments in /opt/conda/lib/python3.7/site-package
s (from ipython<8.0,>=7.23.1->ipykernel>=4.5.2->datalab) (2.10.0)
Requirement already satisfied: pickleshare in /opt/conda/lib/python3.7/site-pack
ages (from ipython<8.0,>=7.23.1->ipykernel>=4.5.2->datalab) (0.7.5)
Requirement already satisfied: pexpect>4.3 in /opt/conda/lib/python3.7/site-pack
ages (from ipython<8.0,>=7.23.1->ipykernel>=4.5.2->datalab) (4.8.0)
Requirement already satisfied: backcall in /opt/conda/lib/python3.7/site-package
s (from ipython<8.0,>=7.23.1->ipykernel>=4.5.2->datalab) (0.2.0)
Requirement already satisfied: prompt-toolkit!=3.0.0,!=3.0.1,<3.1.0,>=2.0.0 in /
opt/conda/lib/python3.7/site-packages (from ipython<8.0,>=7.23.1->ipykernel>=4.
5.2->datalab) (3.0.19)
Requirement already satisfied: decorator in /opt/conda/lib/python3.7/site-packag
es (from ipython<8.0,>=7.23.1->ipykernel>=4.5.2->datalab) (5.0.9)
Requirement already satisfied: parso<0.9.0,>=0.8.0 in /opt/conda/lib/python3.7/s
ite-packages (from jedi>=0.16->ipython<8.0,>=7.23.1->ipykernel>=4.5.2->datalab)
(0.8.2)
Requirement already satisfied: MarkupSafe>=0.23 in /opt/conda/lib/python3.7/site
-packages (from jinja2>=2.8->pandas-profiling==1.4.0->datalab) (1.1.1)
Requirement already satisfied: pyrsistent>=0.14.0 in /opt/conda/lib/python3.7/si
te-packages (from jsonschema>=2.6.0->datalab) (0.17.3)
Requirement already satisfied: attrs>=17.4.0 in /opt/conda/lib/python3.7/site-pa
ckages (from jsonschema>=2.6.0->datalab) (21.2.0)
Requirement already satisfied: jupyter-core>=4.6.0 in /opt/conda/lib/python3.7/s
ite-packages (from jupyter-client<8.0->ipykernel>=4.5.2->datalab) (4.7.1)
Requirement already satisfied: pyzmq>=13 in /opt/conda/lib/python3.7/site-packag
es (from jupyter-client<8.0->ipykernel>=4.5.2->datalab) (22.2.1)
Requirement already satisfied: kiwisolver>=1.0.1 in /opt/conda/lib/python3.7/sit
e-packages (from matplotlib>=1.4->pandas-profiling==1.4.0->datalab) (1.3.1)
Requirement already satisfied: cycler>=0.10 in /opt/conda/lib/python3.7/site-pac
kages (from matplotlib>=1.4->pandas-profiling==1.4.0->datalab) (0.10.0)
Requirement already satisfied: pillow>=6.2.0 in /opt/conda/lib/python3.7/site-pa
ckages (from matplotlib>=1.4->pandas-profiling==1.4.0->datalab) (8.3.1)
Requirement already satisfied: numpy>=1.16 in /opt/conda/lib/python3.7/site-pack
ages (from matplotlib>=1.4->pandas-profiling==1.4.0->datalab) (1.19.5)
Requirement already satisfied: pbr>=0.11 in /opt/conda/lib/python3.7/site-packag
es (from mock>=2.0.0->datalab) (5.6.0)
Requirement already satisfied: pyasn1>=0.1.7 in /opt/conda/lib/python3.7/site-pa
ckages (from oauth2client>=2.2.0->datalab) (0.4.8)
Requirement already satisfied: ptyprocess>=0.5 in /opt/conda/lib/python3.7/site-
packages (from pexpect>4.3->ipython<8.0,>=7.23.1->ipykernel>=4.5.2->datalab) (0.
7.0)
Requirement already satisfied: tenacity>=6.2.0 in /opt/conda/lib/python3.7/site-
packages (from plotly>=1.12.5->datalab) (8.0.1)
Requirement already satisfied: wcwidth in /opt/conda/lib/python3.7/site-packages
(from prompt-toolkit!=3.0.0,!=3.0.1,<3.1.0,>=2.0.0->ipython<8.0,>=7.23.1->ipyker
nel>=4.5.2->datalab) (0.2.5)
Requirement already satisfied: chardet<5,>=3.0.2 in /opt/conda/lib/python3.7/sit
```

```
e-packages (from requests>=2.9.1->datalab) (4.0.0)
Requirement already satisfied: idna<3,>=2.5 in /opt/conda/lib/python3.7/site-pac
kages (from requests>=2.9.1->datalab) (2.10)
Requirement already satisfied: certifi>=2017.4.17 in /opt/conda/lib/python3.7/si
te-packages (from requests>=2.9.1->datalab) (2021.5.30)
Requirement already satisfied: tifffile>=2019.7.26 in /opt/conda/lib/python3.7/s
ite-packages (from scikit-image>=0.13.0->datalab) (2021.8.8)
Requirement already satisfied: imageio>=2.3.0 in /opt/conda/lib/python3.7/site-p
ackages (from scikit-image>=0.13.0->datalab) (2.9.0)
Requirement already satisfied: scipy>=1.0.1 in /opt/conda/lib/python3.7/site-pac
kages (from scikit-image>=0.13.0->datalab) (1.7.1)
Requirement already satisfied: PyWavelets>=1.1.1 in /opt/conda/lib/python3.7/sit
e-packages (from scikit-image>=0.13.0->datalab) (1.1.1)
Requirement already satisfied: networkx>=2.0 in /opt/conda/lib/python3.7/site-pa
ckages (from scikit-image>=0.13.0->datalab) (2.5)
Requirement already satisfied: joblib>=0.11 in /opt/conda/lib/python3.7/site-pac
kages (from scikit-learn>=0.18.2->datalab) (1.0.1)
Requirement already satisfied: threadpoolctl>=2.0.0 in /opt/conda/lib/python3.7/
site-packages (from scikit-learn>=0.18.2->datalab) (2.2.0)
Requirement already satisfied: ipython-genutils in /opt/conda/lib/python3.7/site
-packages (from traitlets<6.0,>=4.1.0->ipykernel>=4.5.2->datalab) (0.2.0)
Note: you may need to restart the kernel to use updated packages.
```

In [3]:
```
# Call Bigquery and examine in dataframe
import google.datalab.bigquery as bq
```

In [4]:
```
df = bq.Query(query + " LIMIT 100").execute().result().to_dataframe()
```

In [5]:
```
df.head(2)
```

Out[5]:

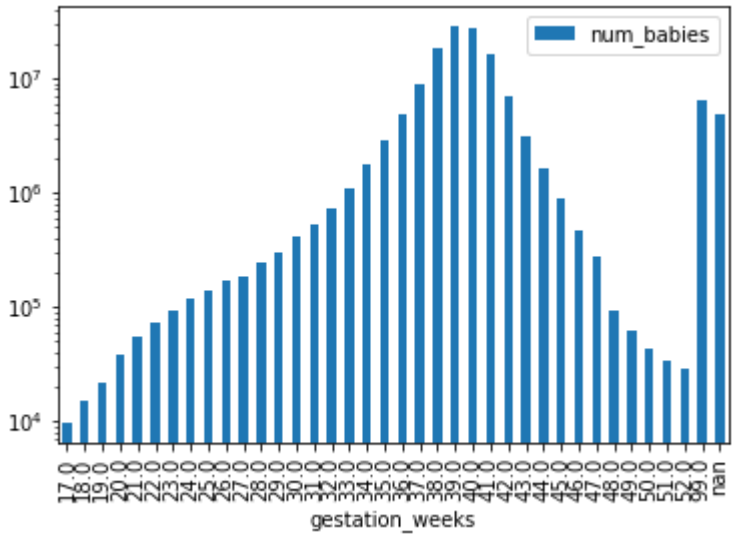
	weight_pounds	is_male	mother_age	plurality	gestation_weeks	f0_
0	7.568469	True	22	1	46	1403073183891835564
1	8.807467	True	39	1	42	1088037545023002395

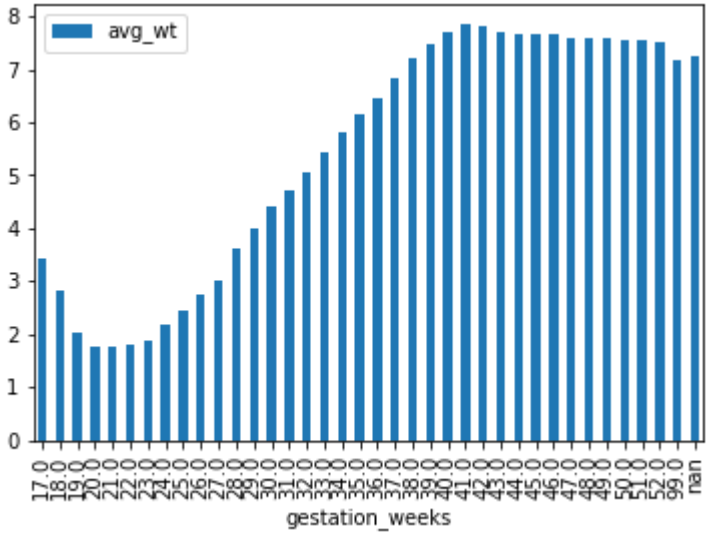
In [6]:
```
# Create function that finds the number of records and the average weight

def get_distinct_values(column_name):
    sql = """
    SELECT
        {0},
        COUNT(1) AS num_babies,
        AVG(weight_pounds) AS avg_wt
    FROM
        publicdata.samples.natality
    WHERE
        year > 200
    GROUP BY
        {0}
    """.format(column_name)
    return bq.Query(sql).execute().result().to_dataframe()
```

In [7]:
```
# Bar plot to use gestation_weeks with avg_wt linear and num_babies
```

```
df = get_distinct_values('gestation_weeks')
df = df.sort_values('gestation_weeks')
df.plot(x='gestation_weeks', y = 'num_babies', logy=True, kind='bar')
df.plot(x='gestation_weeks',y='avg_wt',kind='bar');
```

```
Updated property [core/project].
```





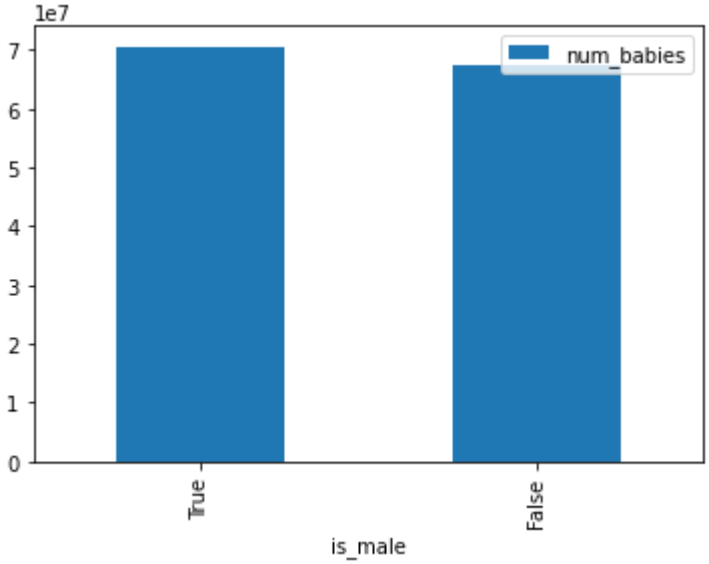
In [8]:

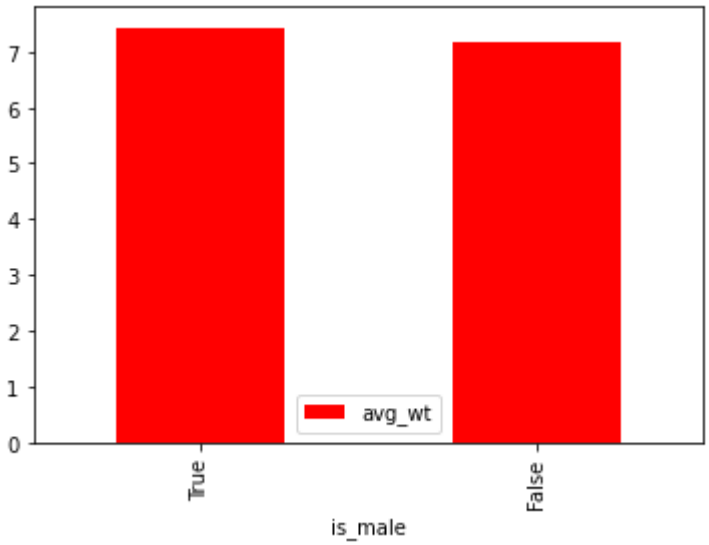
```
# Bar plot to see is_male with avg_wt linear and num_babies logrithmic
df = get_distinct_values('is_male')
df.plot(x='is_male',y='num_babies', kind='bar');
df.plot(x='is_male', y='avg_wt', kind='bar', color='red')
```

```
Updated property [core/project].
```

Out[8]:

```
<AxesSubplot:xlabel='is_male'>
```

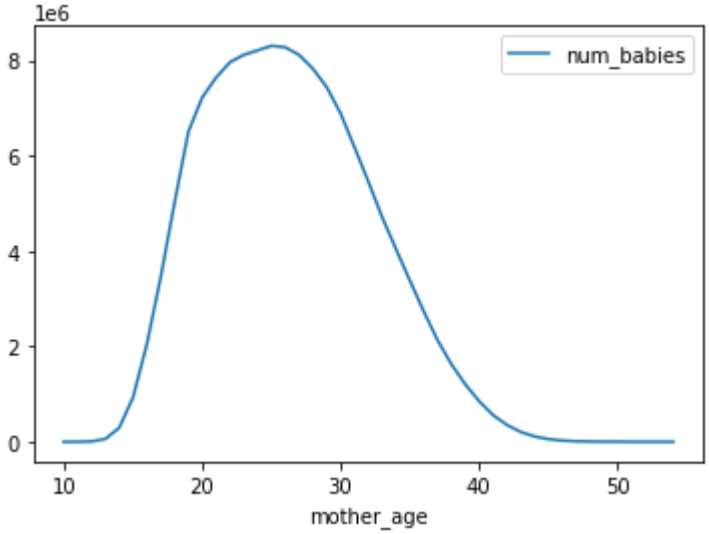


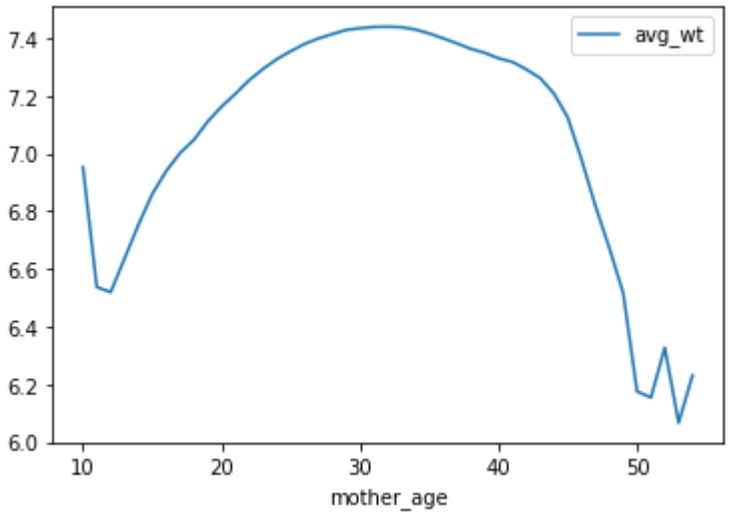


In [9]:
```
# Line plots to see mother_age with average weight(avg_wt) linear and num_babies
df = get_distinct_values('mother_age')
df = df.sort_values('mother_age')
df.plot(x='mother_age', y='num_babies')
df.plot(x='mother_age', y='avg_wt')
```

```
Updated property [core/project].
```

Out[9]:
```
<AxesSubplot:xlabel='mother_age'>
```

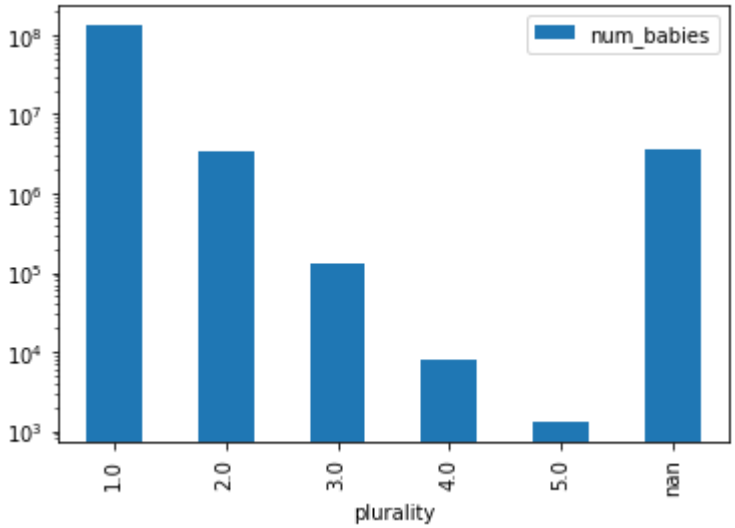


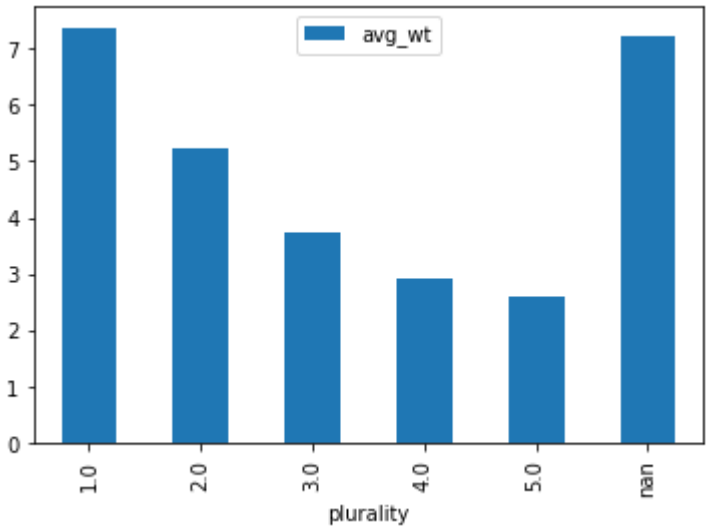


In [10]:
```
# Bar plot to see plurality (singleton, twins, etc,) with avg_wt linear
df = get_distinct_values('plurality')
df = df.sort_values('plurality')
df.plot(x='plurality', y='num_babies', logy=True,kind='bar');
df.plot(x='plurality',y='avg_wt',kind='bar')
```

```
Updated property [core/project].
```

Out[10]:
```
<AxesSubplot:xlabel='plurality'>
```

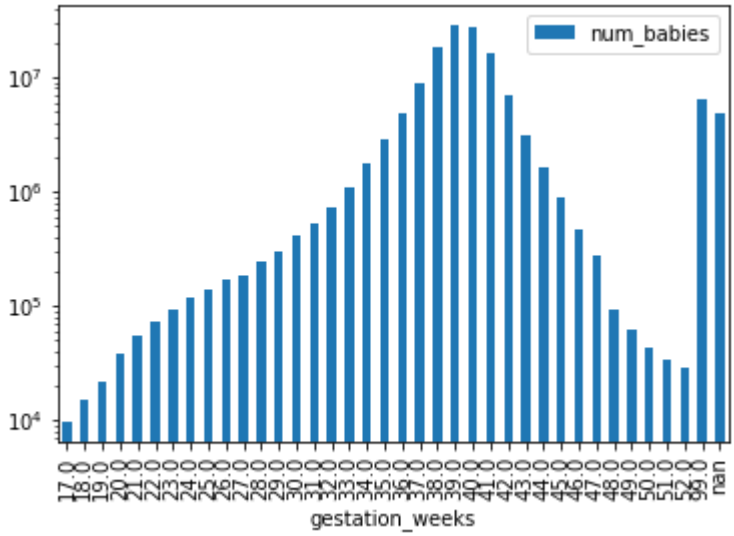


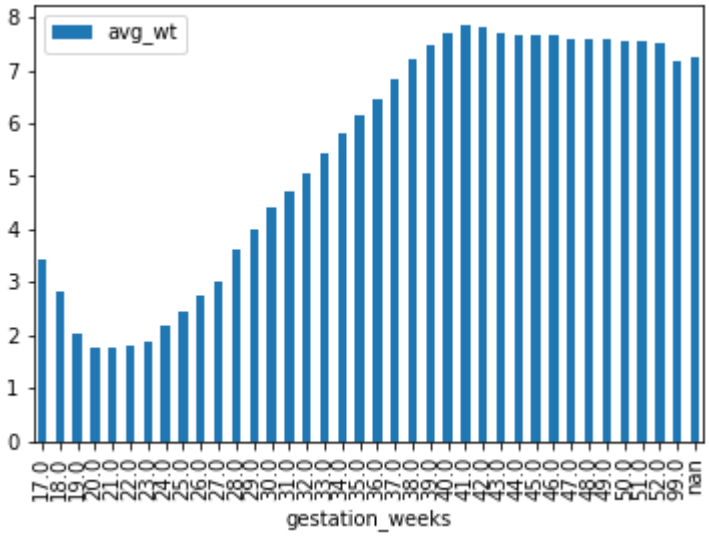


In [11]:
```
# Bar plot to see getation_weeks with avg_wt linear and num_babies

df = get_distinct_values('gestation_weeks')
df=df.sort_values('gestation_weeks')
df.plot(x='gestation_weeks',y='num_babies', logy=True, kind='bar')
df.plot(x='gestation_weeks', y='avg_wt', kind='bar')
```

```
Updated property [core/project].
```

Out[11]:
```
<AxesSubplot:xlabel='gestation_weeks'>
```





All these factors seem to play a part in the baby's weight. Male babies are heavier on average than female babies. Teenaged and older moms tend to have lower-weight babies. Twins, triplets, etc. are lower weight than single births. Preemies weigh in lower as do babies born to single moms. In addition, it is important to check whether you have enough data (number of babies) for each input value. Otherwise, the model prediction against input values that doesn't have enough data may not be reliable.

In [ ]: