In [1]:

```
%load_ext autoreload
%autoreload 2

import os
import sys
# Load Spark python requirements into current scope
# TODO update directory:
spark_home = "/usr/local/lib/python3.10/dist-packages/pyspark/"

sys.path.insert(0, os.path.join(spark_home, 'python/lib/py4j-0.10.9.5-src.zip'))
sys.path.insert(0, os.path.join(spark_home, 'python'))
with open(os.path.join(spark_home, 'python/pyspark/shell.py')) as f:
    code = compile(f.read(), os.path.join(spark_home, 'python/pyspark/shell.py'), 'exec'
    exec(code)
import pandas as pd
print(sc) # Details about spark
```

```
22/12/29 03:15:06 WARN Utils: Your hostname, shameer-laptop resolves to a loopback addre
ss: 127.0.1.1; using 10.186.4.184 instead (on interface wlp2s0)
22/12/29 03:15:06 WARN Utils: Set SPARK_LOCAL_IP if you need to bind to another address
```

```
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLev
el).
```

```
Welcome to
      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /__ / .__/\_,_/_/ /_/\_\   version 3.3.1
      /_/

Using Python version 3.10.6 (main, Nov 14 2022 16:10:14)
Spark context Web UI available at http://pc-4-184.customer.ask4.lan:4040
Spark context available as 'sc' (master = local[*], app id = local-1672265709609).
SparkSession available as 'spark'.
<SparkContext master=local[*] appName=pyspark-shell>
```

In [2]:

```
from pyspark.sql.functions import col

df = spark.read.json("btd2.json")
df = df.withColumn("Duration", col("Duration").cast("double"))
df.printSchema()
df.show()
```

```
root
 |-- Bike #: string (nullable = true)
 |-- Duration: double (nullable = true)
 |-- End Date: string (nullable = true)
 |-- End Station: string (nullable = true)
 |-- End Terminal: string (nullable = true)
 |-- Start Date: string (nullable = true)
 |-- Start Station: string (nullable = true)
 |-- Start Terminal: string (nullable = true)
 |-- Subscription Type: string (nullable = true)
 |-- Trip ID: string (nullable = true)
 |-- Zip Code: string (nullable = true)

+------+--------+-------------+--------------------+------------+-------------+---------
-----------+--------------+-----------------+-------+--------+
|Bike #|Duration|     End Date|         End Station|End Terminal|   Start Date|       St
art Station|Start Terminal|Subscription Type|Trip ID|Zip Code|
+------+--------+-------------+--------------------+------------+-------------+---------
-----------+--------------+-----------------+-------+--------+
|   520|    63.0|8/29/13 14:14|South Van Ness at...|          66|8/29/13 14:13|South Van
Ness at...|            66|       Subscriber|   4576|   94127|
|   661|    70.0|8/29/13 14:43|  San Jose City Hall|          10|8/29/13 14:42|  San Jos
e City Hall|            10|       Subscriber|   4607|   95138|
|    48|    71.0|8/29/13 10:17|Mountain View Cit...|          27|8/29/13 10:16|Mountain 
View Cit...|            27|       Subscriber|   4130|   97214|
|    26|    77.0|8/29/13 11:30|  San Jose City Hall|          10|8/29/13 11:29|  San Jos
e City Hall|            10|       Subscriber|   4251|   95060|
|   319|    83.0|8/29/13 12:04|      Market at 10th|          67|8/29/13 12:02|South Van
Ness at...|            66|       Subscriber|   4299|   94103|
|   527|   103.0|8/29/13 18:56| Golden Gate at Polk|          59|8/29/13 18:54| Golden G
ate at Polk|            59|       Subscriber|   4927|   94109|
|   679|   109.0|8/29/13 13:27|    Adobe on Almaden|           5|8/29/13 13:25|Santa Cla
ra at Al...|             4|       Subscriber|   4500|   95112|
|   687|   111.0|8/29/13 14:04| San Salvador at 1st|           8|8/29/13 14:02| San Salv
ador at 1st|             8|       Subscriber|   4563|   95112|
|   553|   113.0|8/29/13 17:03|South Van Ness at...|          66|8/29/13 17:01|South Van
Ness at...|            66|       Subscriber|   4760|   94103|
|   107|   114.0|8/29/13 11:35|         MLK Library|          11|8/29/13 11:33|  San Jos
e City Hall|            10|       Subscriber|   4258|   95060|
|   368|   125.0|8/29/13 13:55|Embarcadero at Br...|          54|8/29/13 13:52|     Spea
r at Folsom|            49|       Subscriber|   4549|   94109|
|    26|   126.0|8/29/13 13:25|Santa Clara at Al...|           4|8/29/13 13:23|    San P
edro Square|             6|       Subscriber|   4498|   95112|
|   140|   129.0|8/29/13 19:35|Mountain View Cal...|          28|8/29/13 19:32|Mountain 
View Cal...|            28|       Subscriber|   4965|   94041|
|   371|   130.0|8/29/13 13:59|   2nd at South Park|          64|8/29/13 13:57|   2nd at
South Park|            64|       Subscriber|   4557|   94122|
|   503|   134.0|8/29/13 12:33|     Beale at Market|          56|8/29/13 12:31|     Clay
at Battery|            41|       Subscriber|   4386|   94109|
|   408|   138.0|8/29/13 16:59|     Post at Kearney|          47|8/29/13 16:57|     Post
at Kearney|            47|       Subscriber|   4749|   94117|
|    26|   141.0|8/29/13 11:27|  San Jose City Hall|          10|8/29/13 11:25|  San Jos
e City Hall|            10|       Subscriber|   4242|   95060|
|   319|   142.0|8/29/13 12:14|      Market at 10th|          67|8/29/13 12:11|      Mar
ket at 10th|            67|       Subscriber|   4329|   94103|
|   564|   142.0|8/29/13 22:24|Harry Bridges Pla...|          50|8/29/13 22:21|   Steuar
t at Market|            74|       Subscriber|   5097|   94115|
|   574|   144.0|8/29/13 22:08|       Market at 4th|          76|8/29/13 22:06|  Powell 
Street BART|            53|       Subscriber|   5084|   94115|
+------+--------+-------------+--------------------+------------+-------------+---------
-----------+--------------+-----------------+-------+--------+
only showing top 20 rows
```

In [3]:

```
import chart_studio.plotly as py
import plotly.graph_objects as go


def plot_histogram(df):
    fig = go.Figure(
        layout=dict(
        title="Duration Histogram",
        yaxis_type="log",
        xaxis_title='duration',
        yaxis_title='count',
        xaxis_range=[0,59.9]
        )
    )
#     fig.show(renderer="notebook")
    # Extracting data from the RDD df column ['col1']
    data = go.Histogram(x=df.toPandas()['Duration'])
    # Plot the data
    fig.add_trace(data)
    fig.show()
```

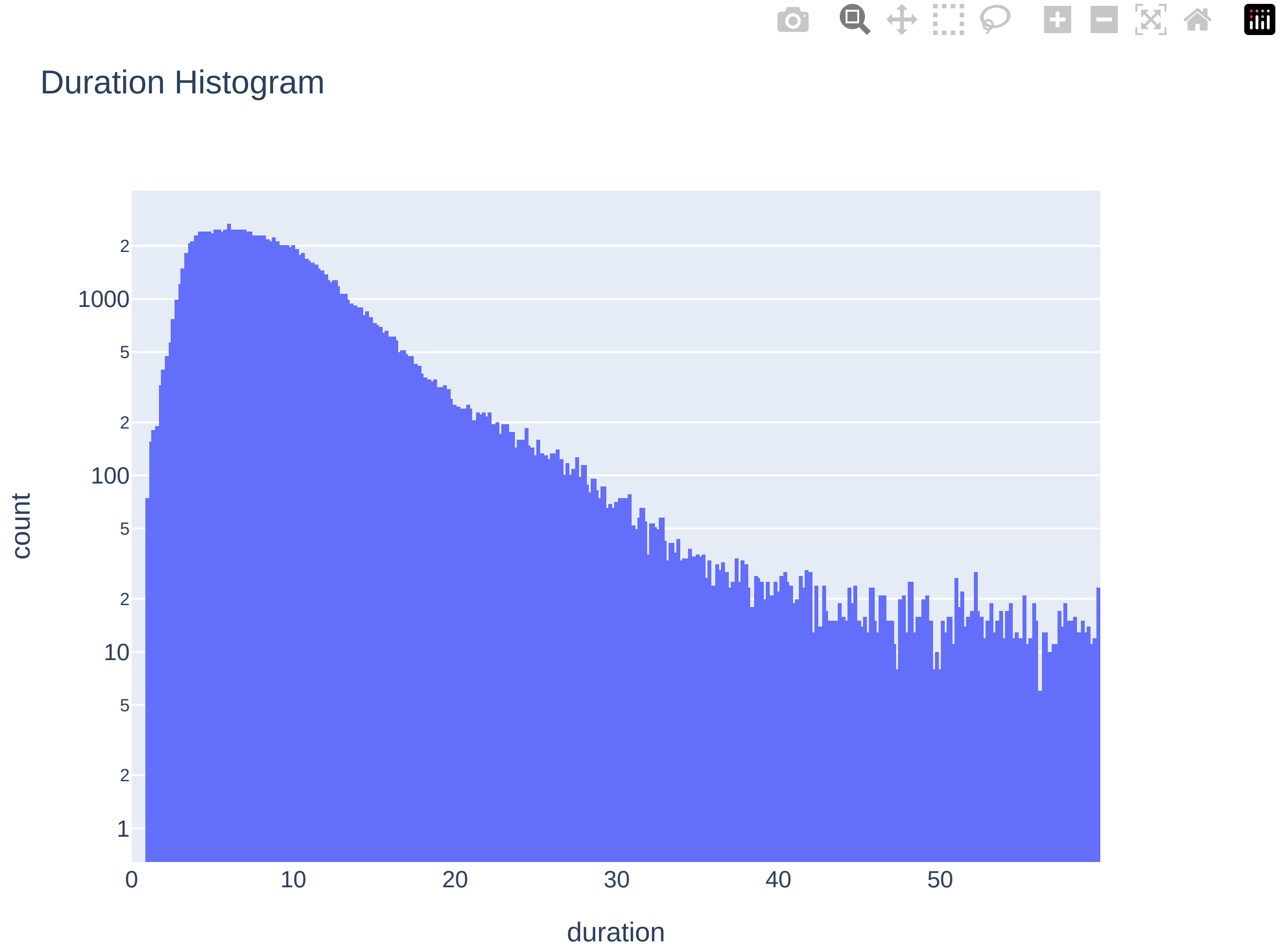
In [4]:

```
# normalizing duration, seconds to min
df = df.withColumn("Duration", col("Duration")/60)
```

In [5]:

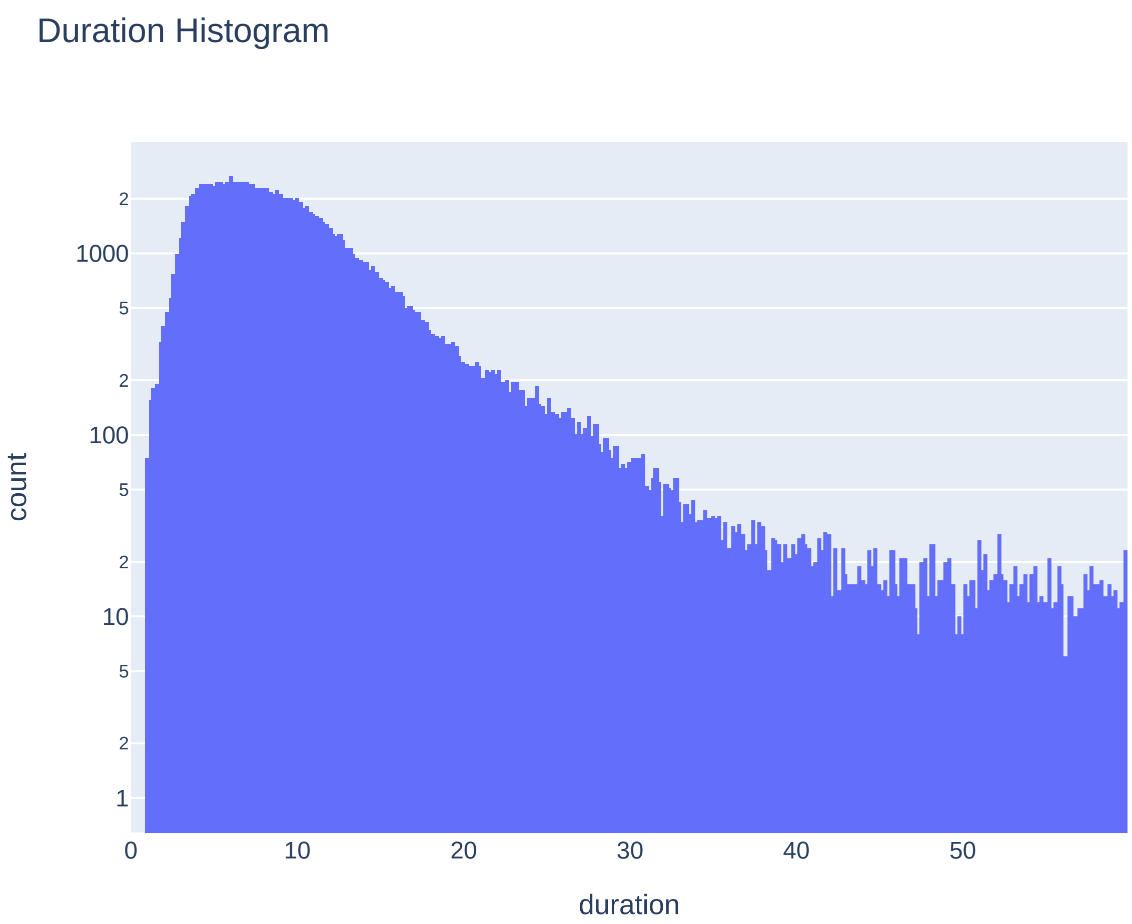
```
## RDD
from pyspark.sql import Row

rdd = df.rdd
duration_df = rdd \
                .filter(lambda r: r['Duration'] < 60.0) \
                .map(lambda r: Row(Duration=r['Duration'])) \
                .toDF()
plot_histogram(duration_df)
```



In [6]:

```
## Dataframe
df.createOrReplaceTempView("bikes")
sql_str = "select Duration from bikes where Duration < 60.0"
sql_duration_df = spark.sql(sql_str)
plot_histogram(sql_duration_df)
```


Duration Histogram
count
2
1000
5
2
100
5
2
10
5
2
1
0
10
20
30
40
50
duration

```
In [ ]:
```