**Top 100 queries:**

--NYC Taxi data

SELECT TOP 1000 \* FROM

    OPENROWSET(

        BULK 'https://azureopendatastorage.blob.core.windows.net/nyctlc/yellow/puYear=\*/puMonth=\*/\*.parquet',

        FORMAT='PARQUET'

    ) AS [nyc]

--Public Holidays data

SELECT TOP 100 \* FROM

    OPENROWSET(

        BULK 'https://azureopendatastorage.blob.core.windows.net/holidaydatacontainer/Processed/\*.parquet',

        FORMAT='PARQUET'

    ) AS [holidays]

--Weather Data

SELECT

    TOP 1000 \*

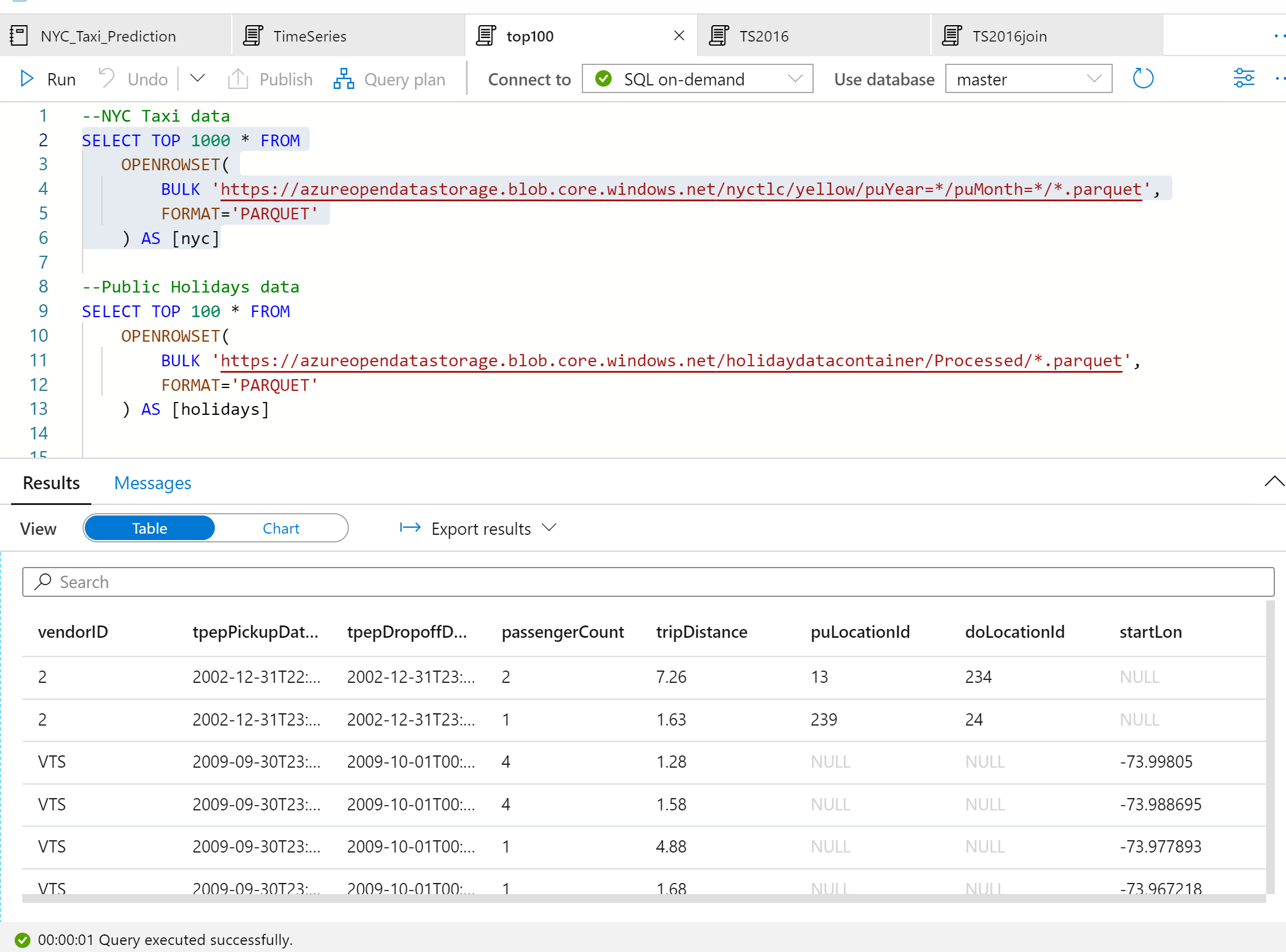
FROM

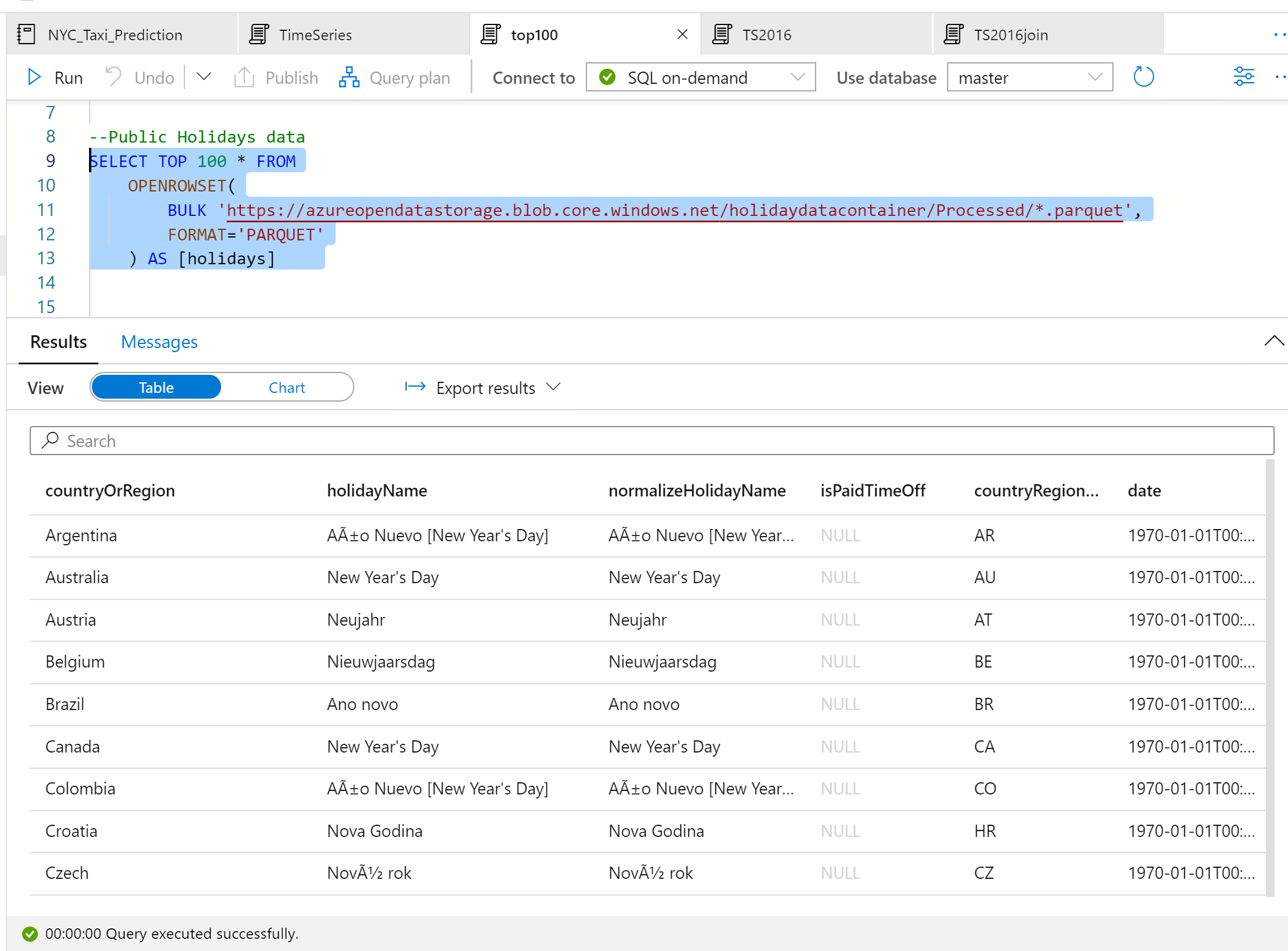
    OPENROWSET(

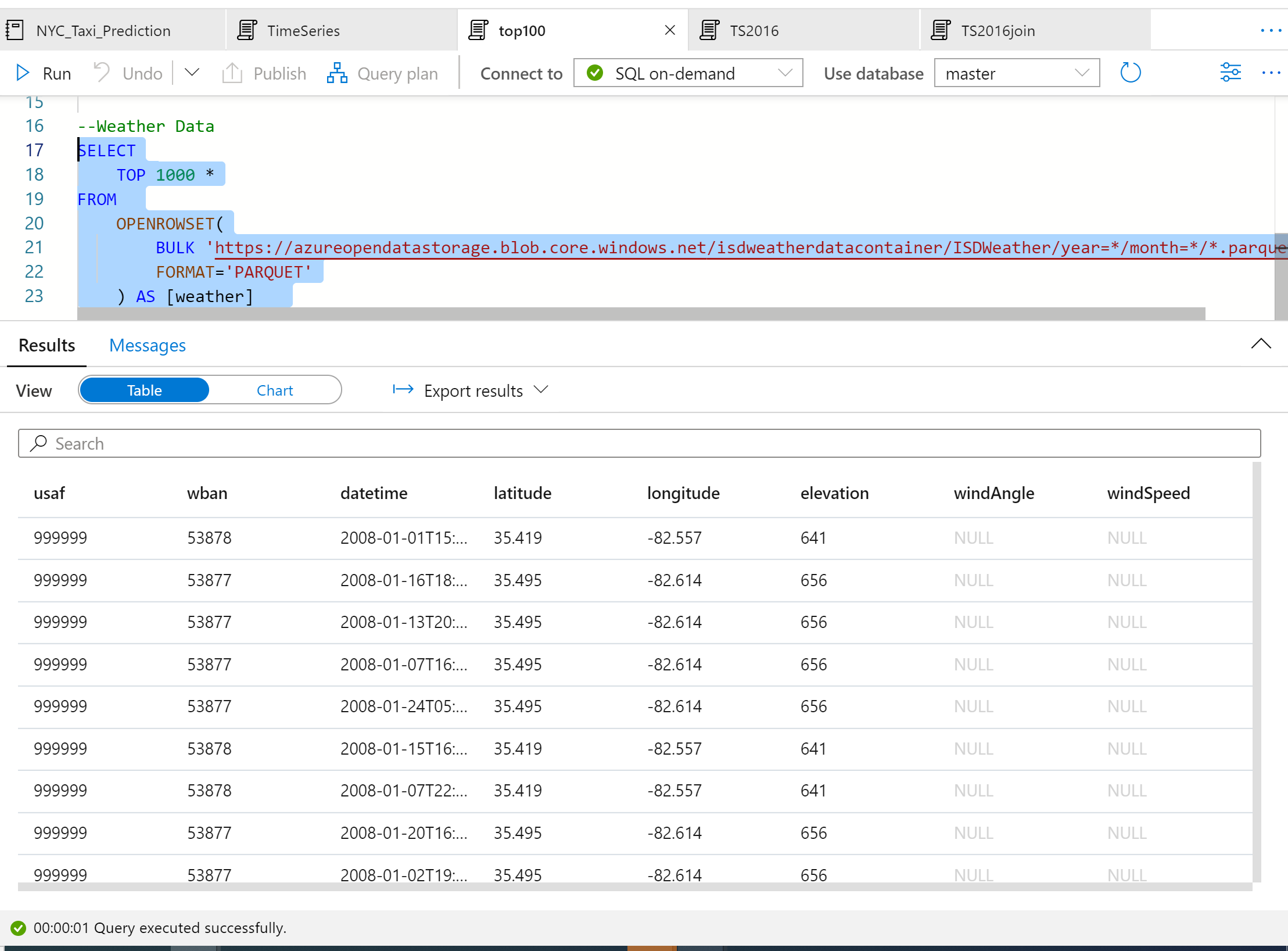
        BULK 'https://azureopendatastorage.blob.core.windows.net/isdweatherdatacontainer/ISDWeather/year=\*/month=\*/\*.parquet',

        FORMAT='PARQUET'

    ) AS [weather]







**Time series, seasonality, and outlier analysis**

### **Summarize the yearly number of taxi rides by using the following query:**

SELECT

    YEAR(tpepPickupDateTime) AS current\_year,

    COUNT(\*) AS rides\_per\_year

FROM

    OPENROWSET(

        BULK 'https://azureopendatastorage.blob.core.windows.net/nyctlc/yellow/puYear=\*/puMonth=\*/\*.parquet',

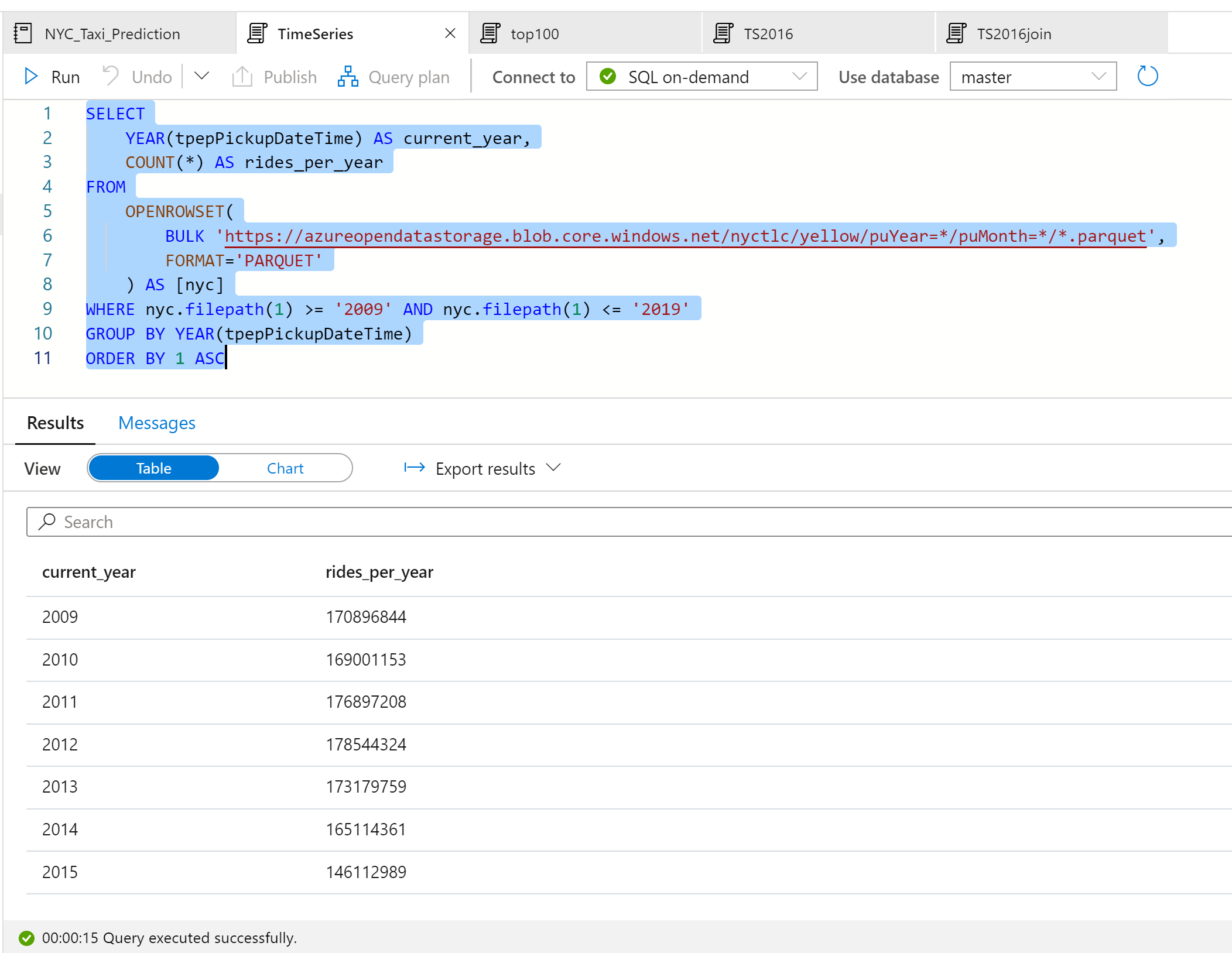
        FORMAT='PARQUET'

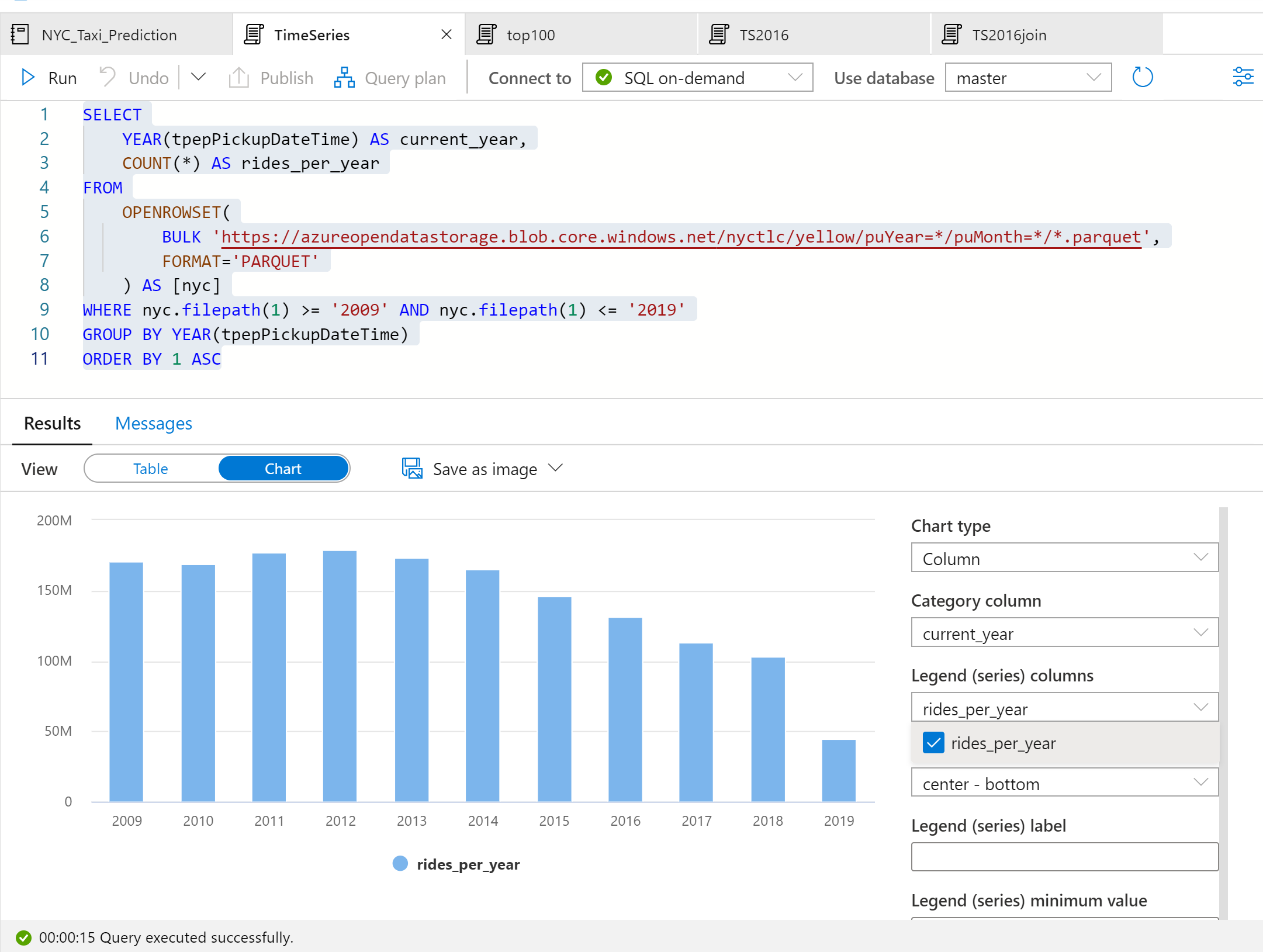
    ) AS [nyc]

WHERE nyc.filepath(1) >= '2009' AND nyc.filepath(1) <= '2019'

GROUP BY YEAR(tpepPickupDateTime)

ORDER BY 1 ASC





Note:  data for 2019 is incomplete. As a result, there's a huge drop in the number of rides for that year.

### **For 2016,the following query returns the daily number of rides during that year:**

SELECT

    CAST([tpepPickupDateTime] AS DATE) AS [current\_day],

    COUNT(\*) as rides\_per\_day

FROM

    OPENROWSET(

        BULK 'https://azureopendatastorage.blob.core.windows.net/nyctlc/yellow/puYear=\*/puMonth=\*/\*.parquet',

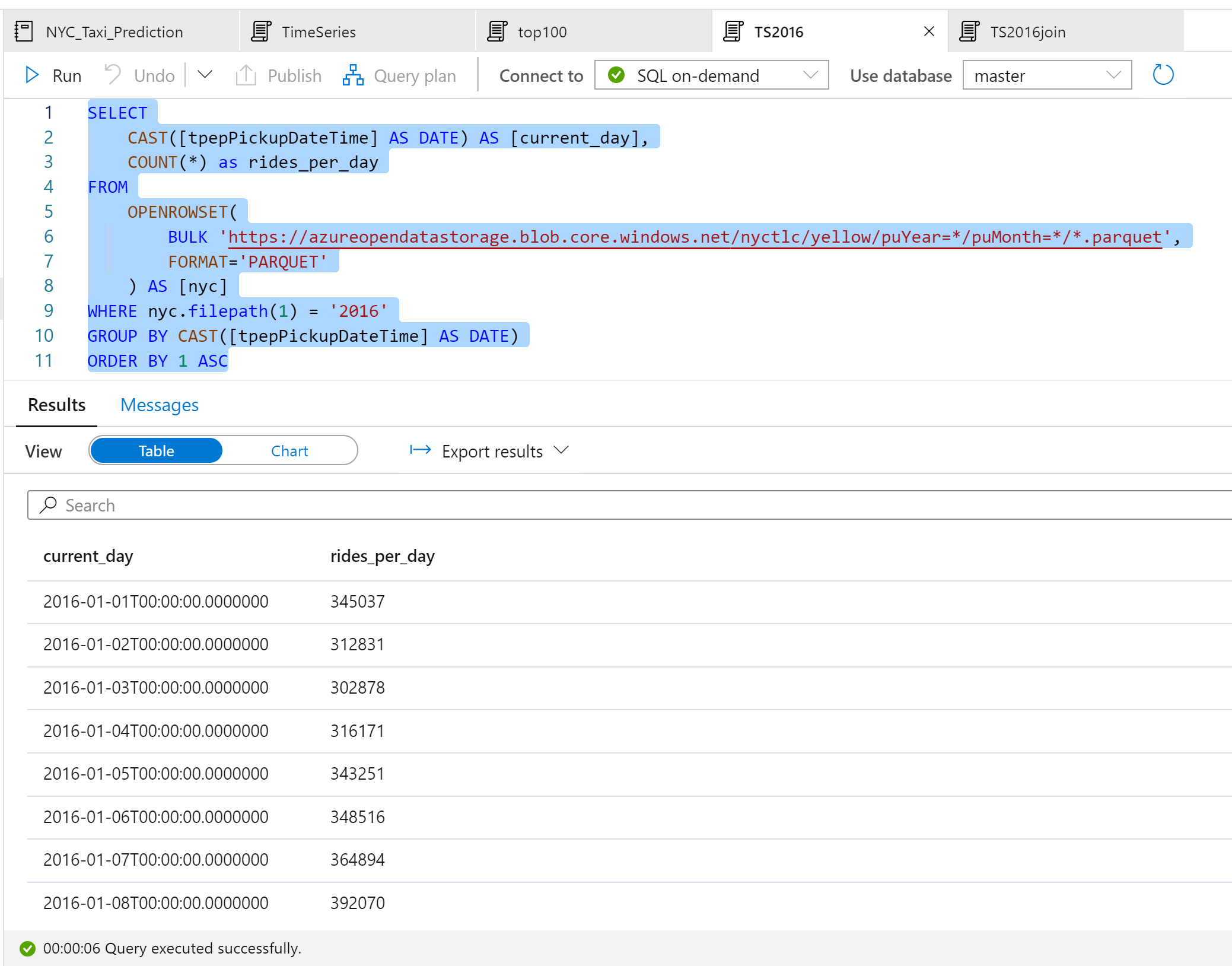
        FORMAT='PARQUET'

    ) AS [nyc]

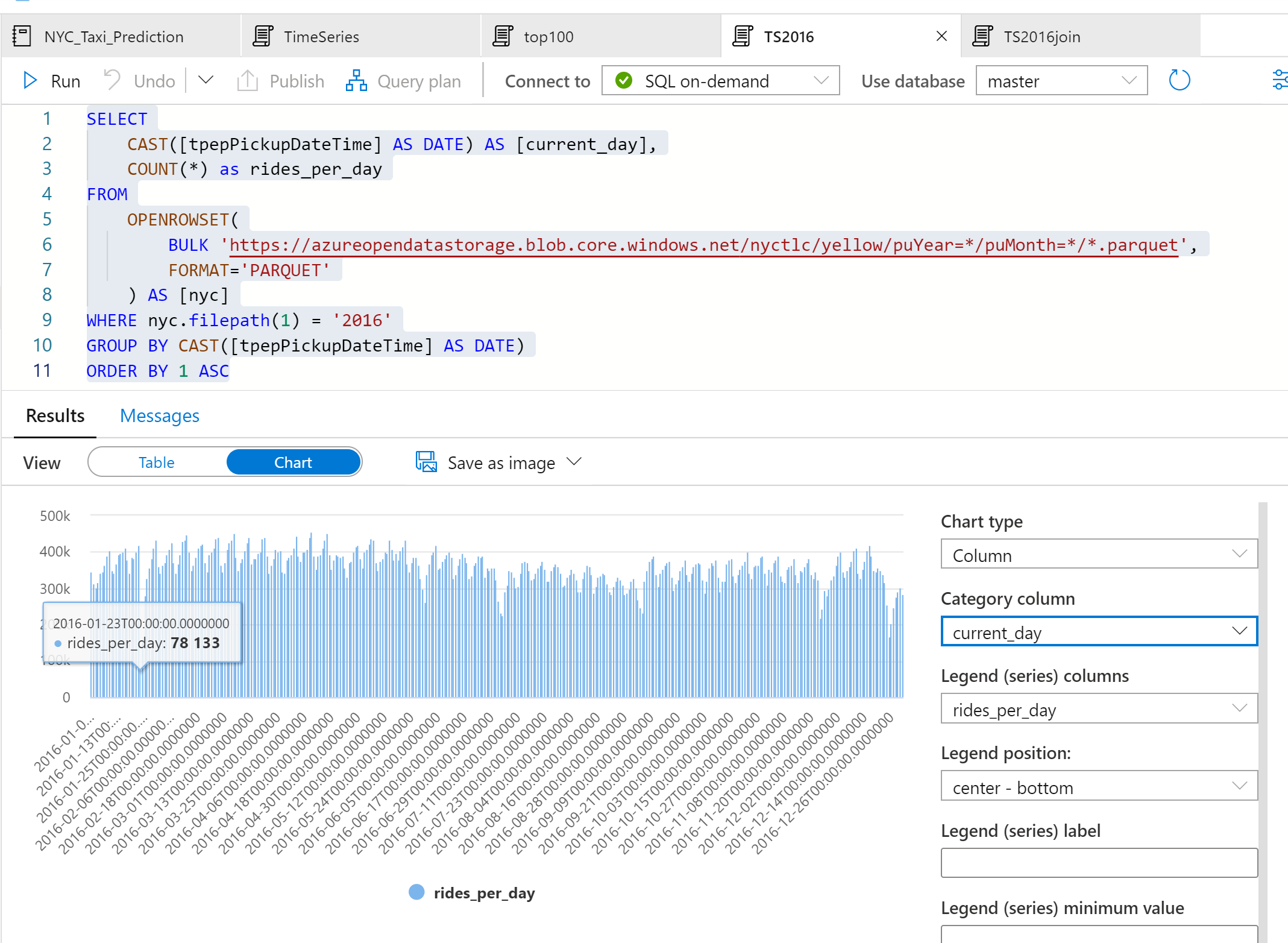
WHERE nyc.filepath(1) = '2016'

GROUP BY CAST([tpepPickupDateTime] AS DATE)

ORDER BY 1 ASC



**Visualize data by plotting the**Column**chart with the**Category**column set to**current\_day**and the**Legend (series)**column set to**rides\_per\_day**.**



From the plot chart, we can see that there's a weekly pattern, with Saturdays as the peak day. During summer months, there are fewer taxi rides because of vacations. There are also some significant drops in the number of taxi rides without a clear pattern of when and why they occur. Next, let us see if the drops correlate with public holidays by joining the NYC Taxi rides dataset with the Public Holidays dataset:

WITH taxi\_rides AS

(

    SELECT

        CAST([tpepPickupDateTime] AS DATE) AS [current\_day],

        COUNT(\*) as rides\_per\_day

    FROM

        OPENROWSET(

            BULK 'https://azureopendatastorage.blob.core.windows.net/nyctlc/yellow/puYear=\*/puMonth=\*/\*.parquet',

            FORMAT='PARQUET'

        ) AS [nyc]

    WHERE nyc.filepath(1) = '2016'

    GROUP BY CAST([tpepPickupDateTime] AS DATE)

),

public\_holidays AS

(

    SELECT

        holidayname as holiday,

        date

    FROM

        OPENROWSET(

            BULK 'https://azureopendatastorage.blob.core.windows.net/holidaydatacontainer/Processed/\*.parquet',

            FORMAT='PARQUET'

        ) AS [holidays]

    WHERE countryorregion = 'United States' AND YEAR(date) = 2016

)

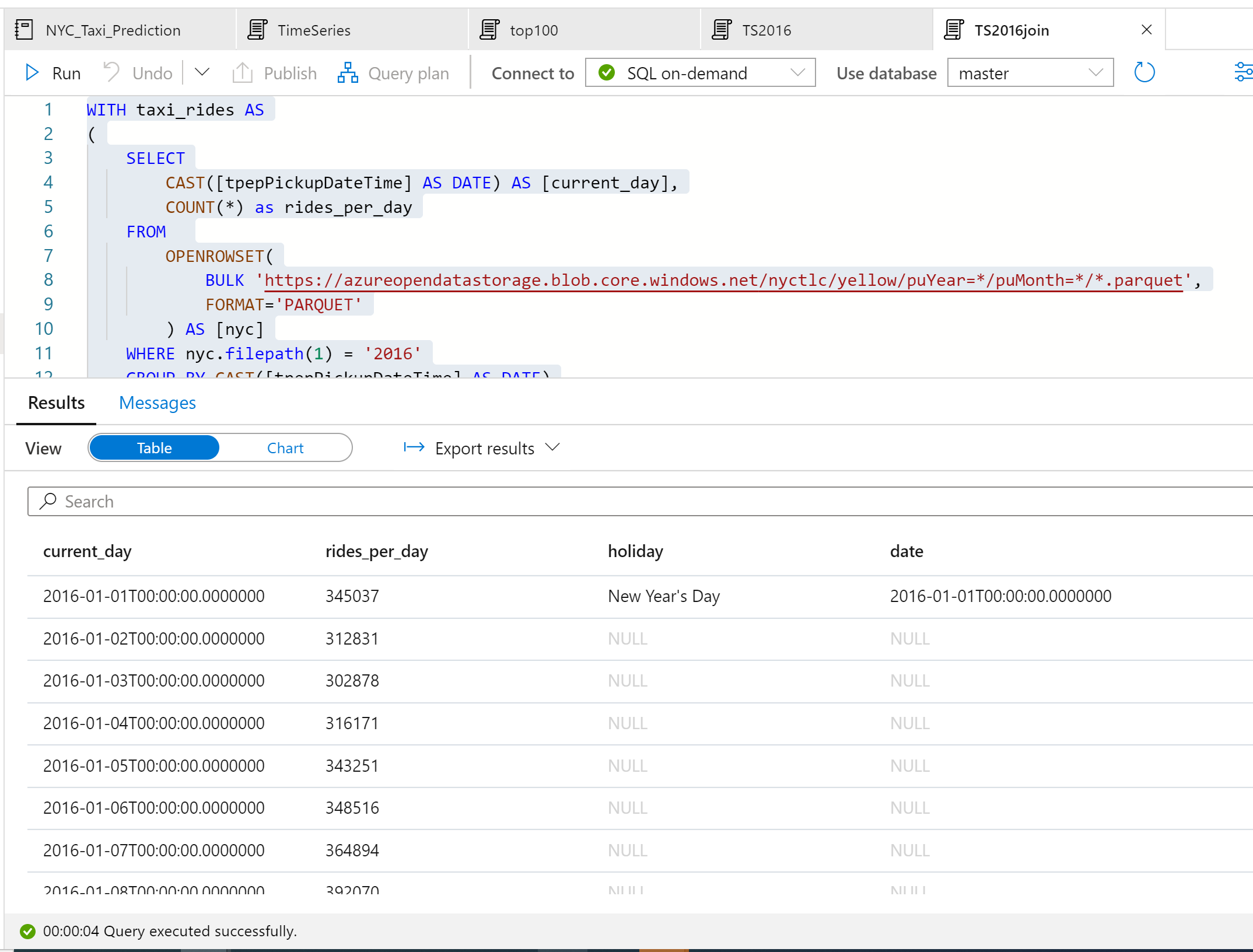
SELECT

\*

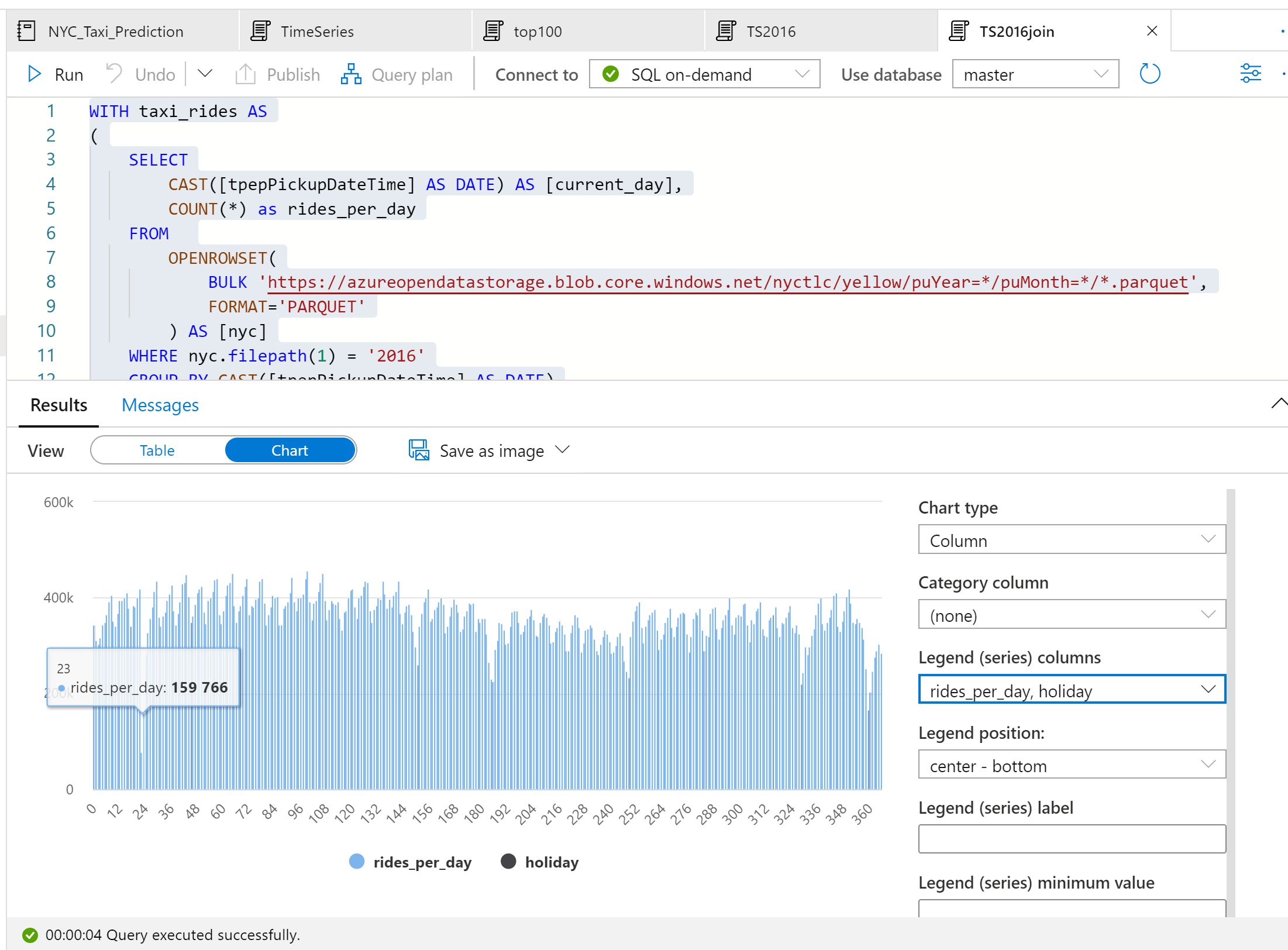
FROM taxi\_rides t

LEFT OUTER JOIN public\_holidays p on t.current\_day = p.date

ORDER BY current\_day ASC



This time, we want to highlight the number of taxi rides during public holidays. For that purpose, we choose **none** for the **Category** column and **rides\_per\_day** and **holiday** as the **Legend (series)** columns.



From the plot chart, we can see that during public holidays the number of taxi rides is lower. There is still one unexplained large drop on January 23. Let us check the weather in NYC on that day by querying the Weather Data dataset:

SELECT

    AVG(windspeed) AS avg\_windspeed,

    MIN(windspeed) AS min\_windspeed,

    MAX(windspeed) AS max\_windspeed,

    AVG(temperature) AS avg\_temperature,

    MIN(temperature) AS min\_temperature,

    MAX(temperature) AS max\_temperature,

    AVG(sealvlpressure) AS avg\_sealvlpressure,

    MIN(sealvlpressure) AS min\_sealvlpressure,

    MAX(sealvlpressure) AS max\_sealvlpressure,

    AVG(precipdepth) AS avg\_precipdepth,

    MIN(precipdepth) AS min\_precipdepth,

    MAX(precipdepth) AS max\_precipdepth,

    AVG(snowdepth) AS avg\_snowdepth,

    MIN(snowdepth) AS min\_snowdepth,

    MAX(snowdepth) AS max\_snowdepth

FROM

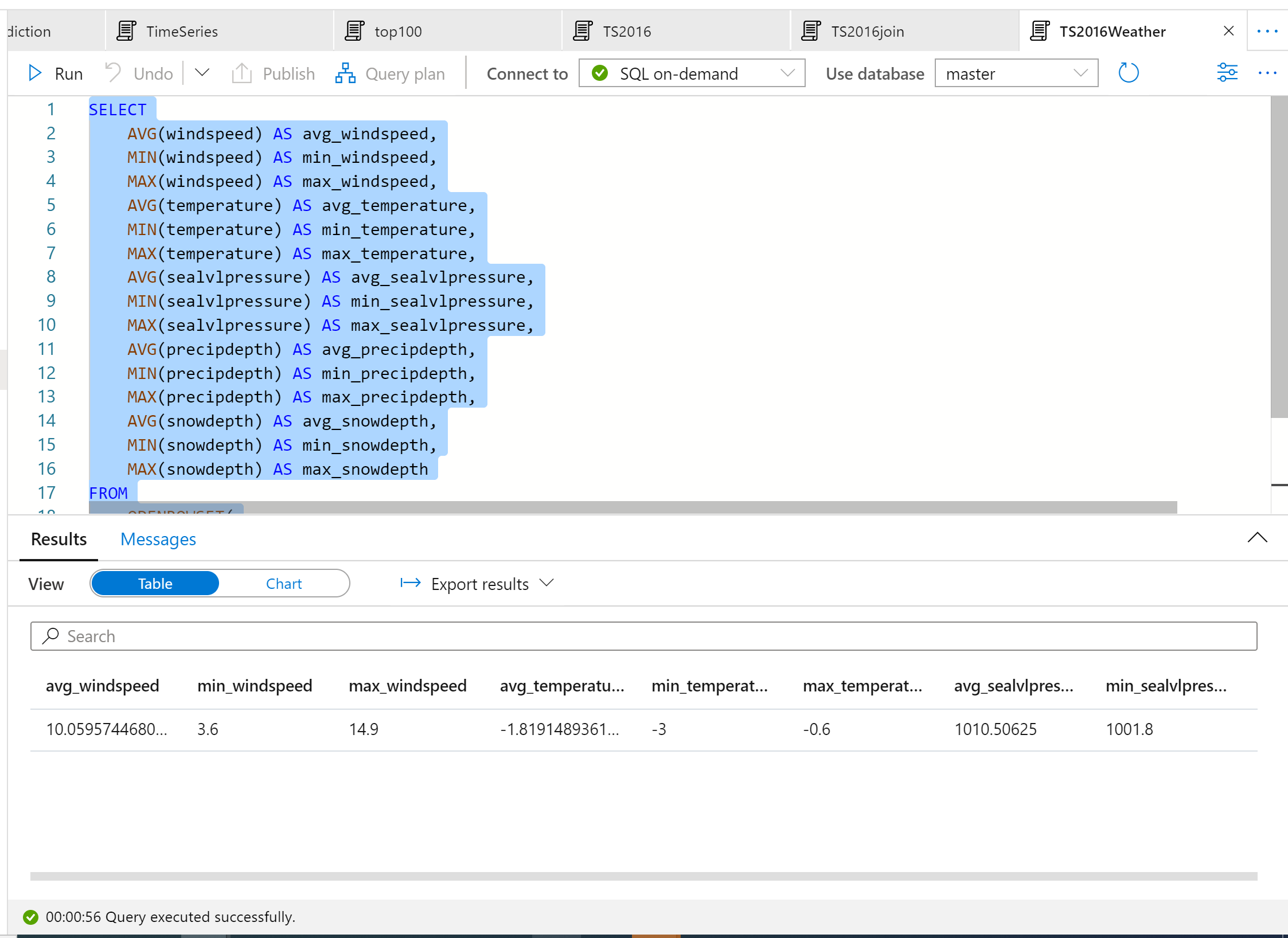
    OPENROWSET(

        BULK 'https://azureopendatastorage.blob.core.windows.net/isdweatherdatacontainer/ISDWeather/year=\*/month=\*/\*.parquet',

        FORMAT='PARQUET'

    ) AS [weather]

WHERE countryorregion = 'US' AND CAST([datetime] AS DATE) = '2016-01-23' AND stationname = 'JOHN F KENNEDY INTERNATIONAL AIRPORT'



The results of the query indicate that the drop in the number of taxi rides occurred because:

* There was a blizzard on that day in NYC with heavy snow (~30 cm).
* It was cold (temperature was below zero degrees Celsius).
* It was windy (~10 m/s).