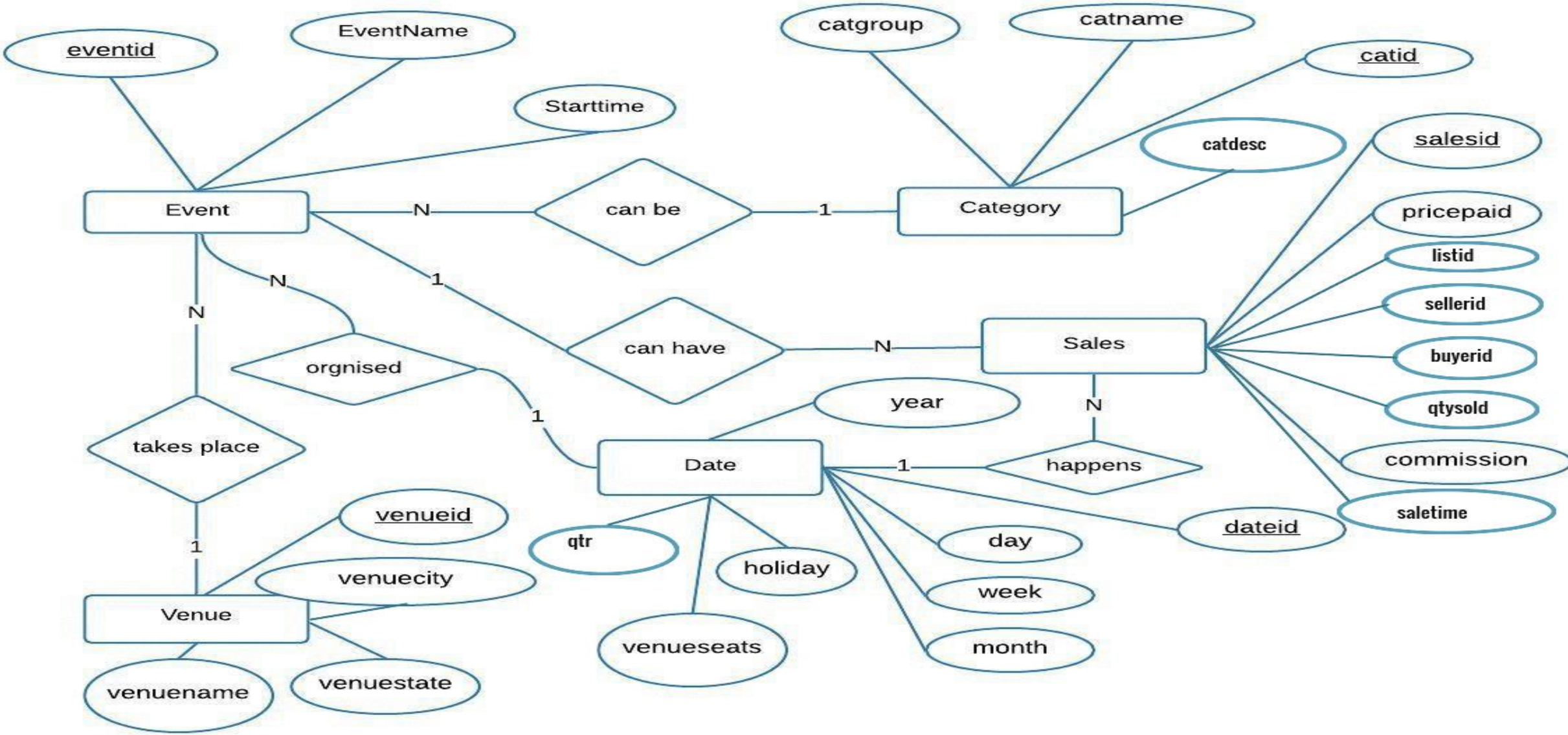
Database System S21
Assignment 10

-Suresh Kumar Choudhary 5504392

-Rohan Chitte 5509596

-Frenny Macwan 5504865

Task 1: ER diagram



Task 1: Relational Schema

- Event (eventid, venueid, catid, dateid, eventname, starttime)
- category(catid, catgroup, catname, catdesc)
- Sales (salesid, eventid, dateid,listid,sellerid, buyerid, qtysold, pricepaid, commission, saletime)
- Venue (venueid,venuename,venuecity,venuestate,venueseats)
- Date (dateid, caldate, day, week, month, qtr, year, holiday)

Objective

- **Analysis of bike sales of year 2008 in the USA**

  - Monthly sales
  - USA state wise sales
  - Events wise sales
  - Category wise sales (ie. Musicals, opera, plays, pop)

Task 2: Cleaning

The dataset contains multiple csv files. So, Pandas library in python was used to read and store the csv files as dataframes.

In [1]:
```
import numpy as np
import pandas as pd
```

In [2]:
```
#Reading csv files into dataframe
category = pd.read_csv(r'/users/rohanchitte/downloads/data-2/category.csv')
date_tab = pd.read_csv(r'/users/rohanchitte/downloads/data-2/date_table.csv')
event = pd.read_csv(r'/users/rohanchitte/downloads/data-2/event.csv')
sales = pd.read_csv(r'/users/rohanchitte/downloads/data-2/sales.csv')
venue = pd.read_csv(r'/users/rohanchitte/downloads/data-2/venue.csv')
```

In [3]:
```
print(venue.isnull().sum())
print(category.isnull().sum())
print(date_tab.isnull().sum())
print(event.isnull().sum())
print(sales.isnull().sum())
```

```
venueid         0
venuename       0
venuecity       0
venuestate      0
venueseats     15
dtype: int64
```

Task 2: Cleaning (Cont.)

Pandas isnull() function was used to find null values and only the column venueseats of venue dataframe had null values.

In [ ]:

```
# 130/200 values are 0, so we will replace null values with 0.
#Moreover, the objective of visualization does not deal with venueseats.
#So, data imputation becomes irrelevant.
venueseats = venue['venueseats']
count = 0
for i in venueseats:
    if i == 0:
        count = count + 1
print(count)  #130

```

In [ ]:

```
#Data Imputation
venue['venueseats'] = venue['venueseats'].fillna(0)
```

Task 2: Import Data

Postgresql Database was used to create a database.
Python package sqlalchemy was used to first establish a connection with the database.
Then, Pandas function to_sql() was used to write and store records of the dataframe to the database.

In [17]:
```
#Importing Data
```

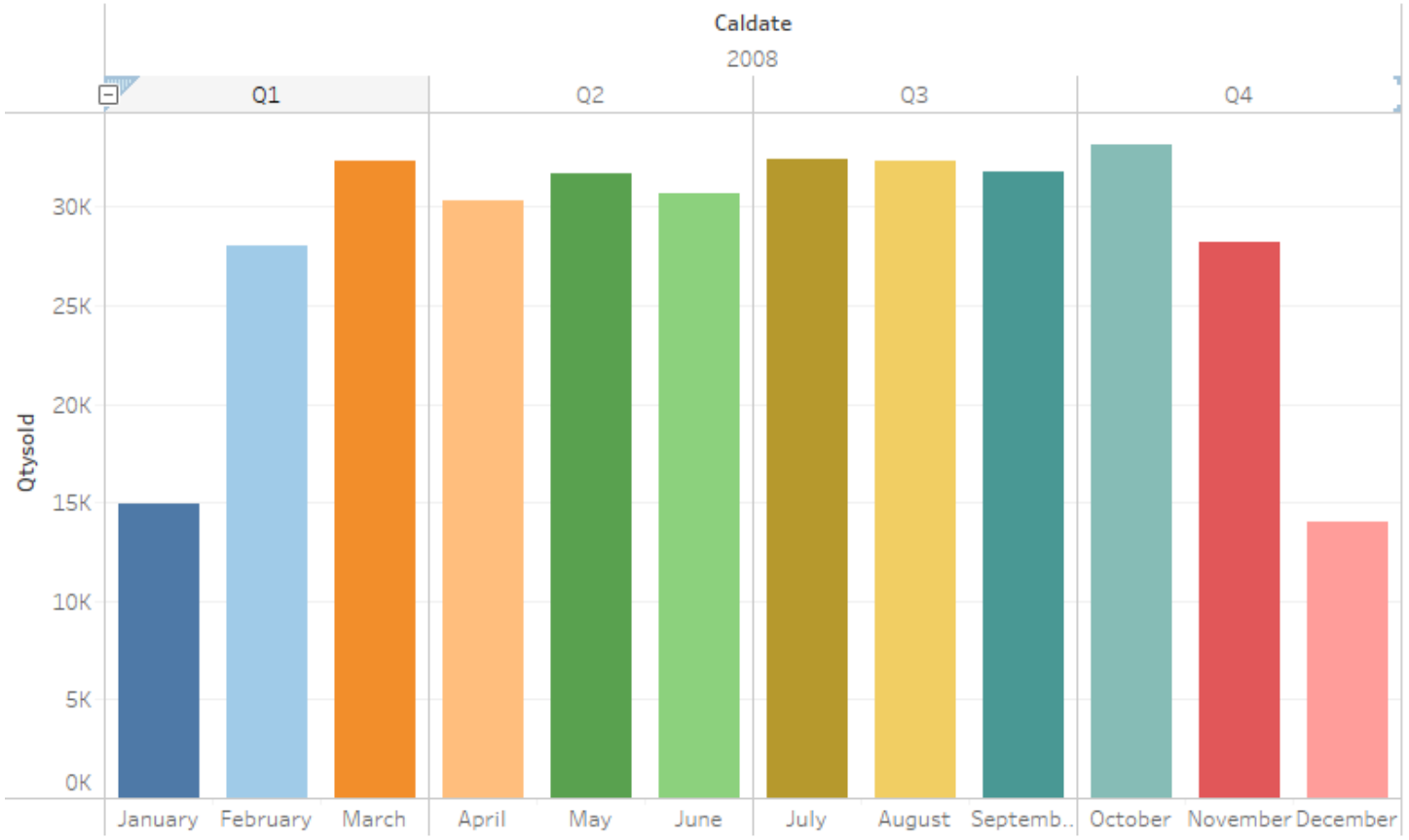
In [26]:
```
from sqlalchemy import create_engine
engine = create_engine('postgresql://:@localhost:5432/databaseproject')
```

In [27]:
```
sales.to_sql('sales', engine)
venue.to_sql('venue', engine)
event.to_sql('event', engine)
date_tab.to_sql('date_table', engine)
category.to_sql('category', engine)
```

Task 3: Visualization of the data set

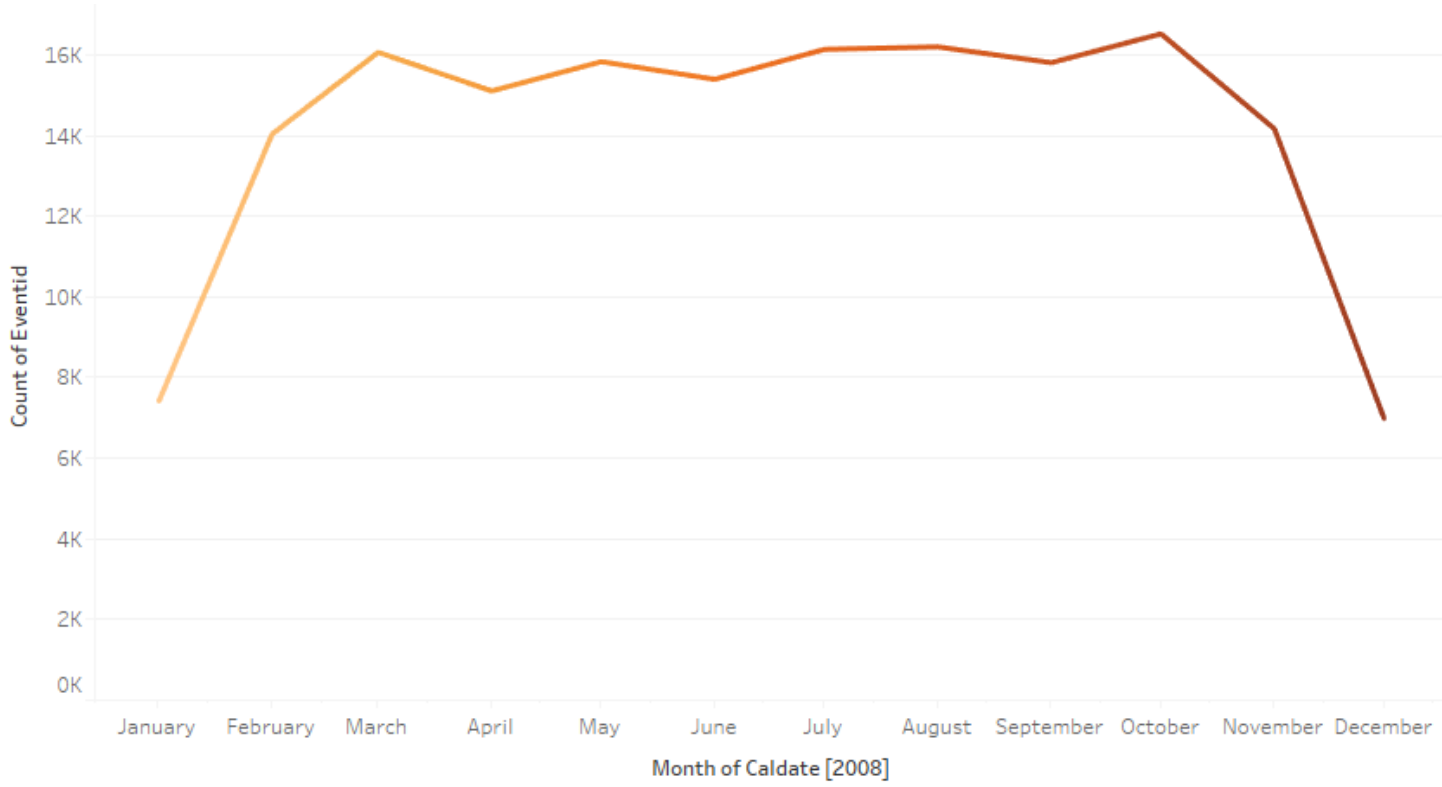
- Analysis of bike sales for year 2008.

What is particularly encouraging is to see how well e-bikes sales have held up in the USA in the challenging market conditions for year 2008. It can be seen from ‘Monthly sales’ graph, the quantity of bikes is sold is quite steady except for two months – January and December.



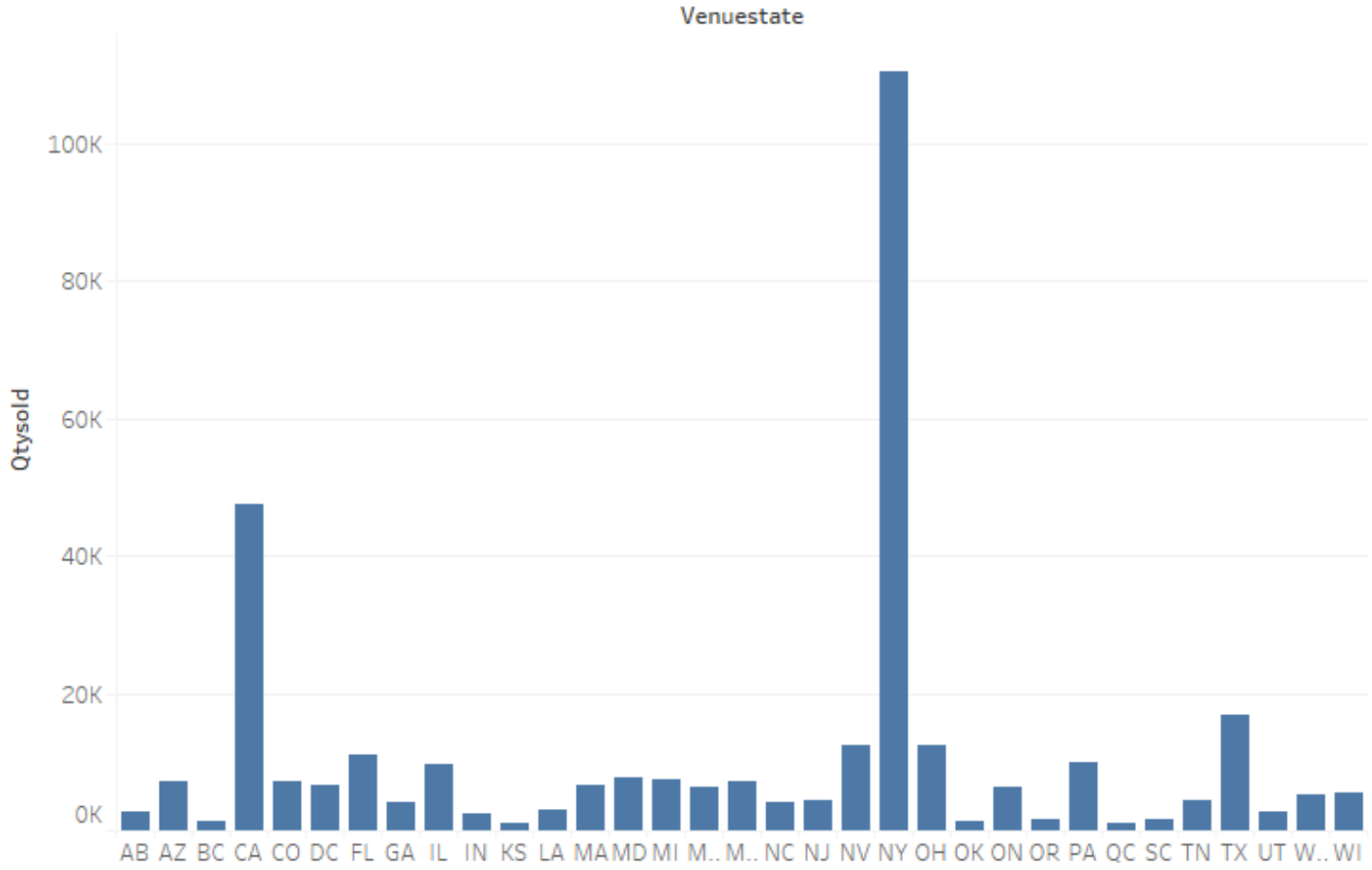
Task 3: Visualization… Cont…

From above graph, number of events happened in January and December months are less compared to other months. That also reflects into quarterly sells. So, we can say that sales of bikes have been increased when more events are happening throughout the year.

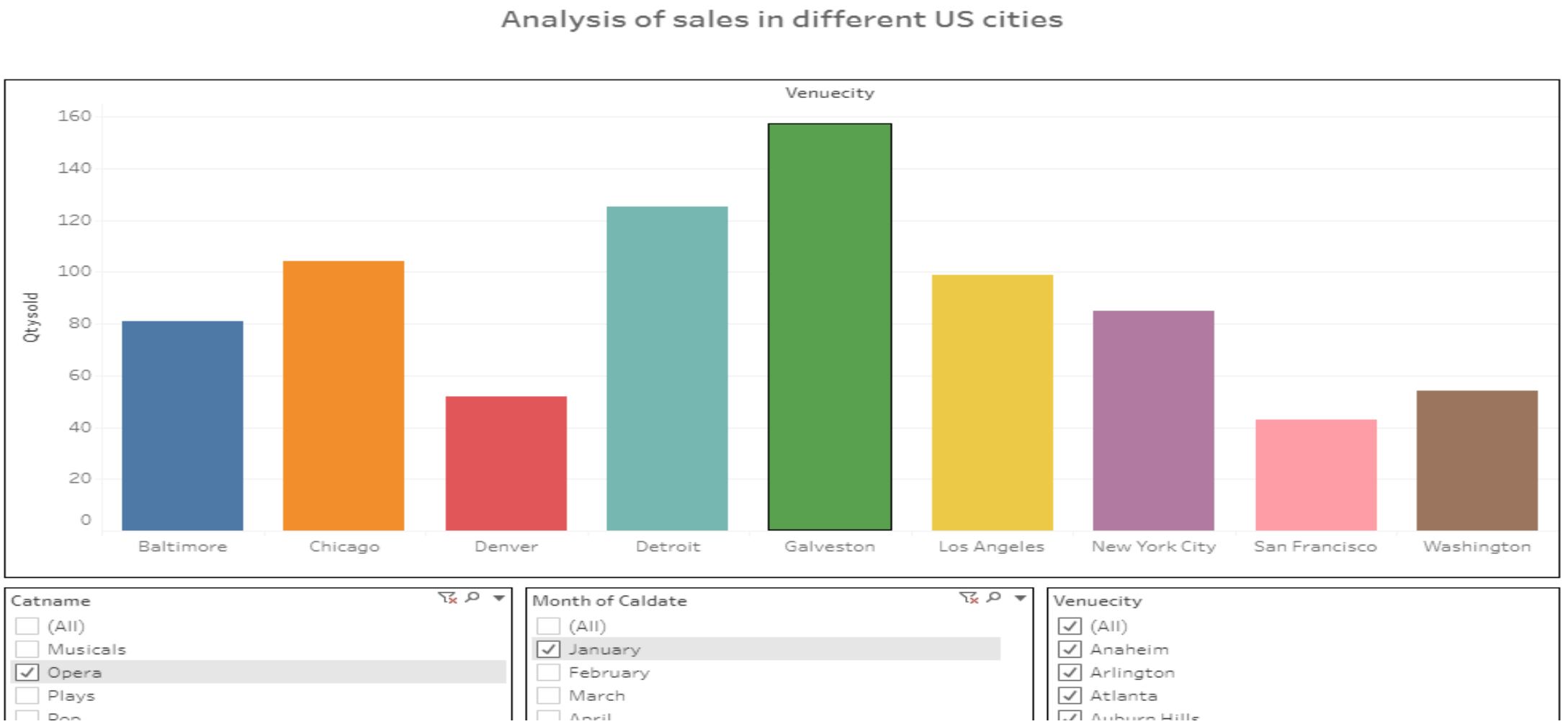


Task 3: Visualization… Cont…

This gives a general overview of the sale of e-bikes in different states of USA. However, it also confirms some known challenges with the different states of the country. We already know NY and CA are main states of the USA, where a greater number of events are taking place. It can be inferred that in these two states, more numbers of bikes are purchased compared to other states.



Task 3: Visualization…
Cont…



- Source code:

- Github link : https://github.com/FrennyMacwan/Database-Project


Thank
you!