visual_data_analysis.ipynb

Markdown ∨   Python 3 ○

Visual Data Analysis of Fraudulent Transactions

```
[1]:  # initial imports
      import pandas as pd
      import calendar
      import plotly.express as px
      import hvplot.pandas
      from sqlalchemy import create_engine
```

```
[2]:  # create a connection to the database
      engine = create_engine("postgresql://postgres:postgres@localhost:5432/fraud_detection")
```

Data Analysis Questions 1

Use `hvPlot` to create a line plot showing a time series from the transactions along all the year for **card holders 2 and 18**. In order to contrast the patterns of both card holders, create a line plot containing both lines. What difference do you observe between the consumption patterns? Does the difference could be a fraudulent transaction? Explain your rationale.

```
[3]:  # loading data for card holder 2 and 18 from the database
      query = """
              SELECT ch.id AS cardholder, t.date AS hour, t.amount
              FROM transaction AS t
              JOIN credit_card AS cc ON cc.card = t.card
              JOIN card_holder AS ch ON ch.id = cc.id_card_holder
              WHERE ch.id in (2, 18)
              ORDER BY hour
              """
      df_question1 = pd.read_sql(query, engine)
```

```
[4]:  # plot for cardholder 2
      data_2 = df_question1[df_question1["cardholder"] == 2]
      data_2_plot = data_2.hvplot.line("hour", "amount", label="Card holder 2", dynamic=False)
      data_2_plot
```
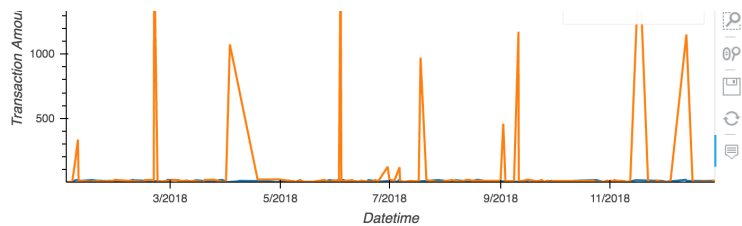
[4]:
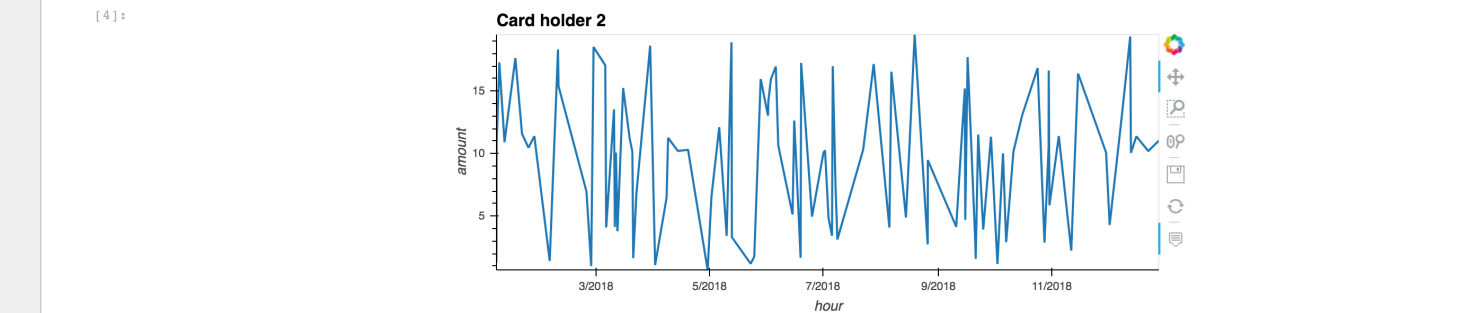

```
[5]:  # plot for cardholder 18
      data_18 = df_question1[df_question1["cardholder"] == 18]
      data_18_plot = data_18.hvplot.line("hour", "amount", label="Card holder 18")
      data_18_plot
```

[5]:
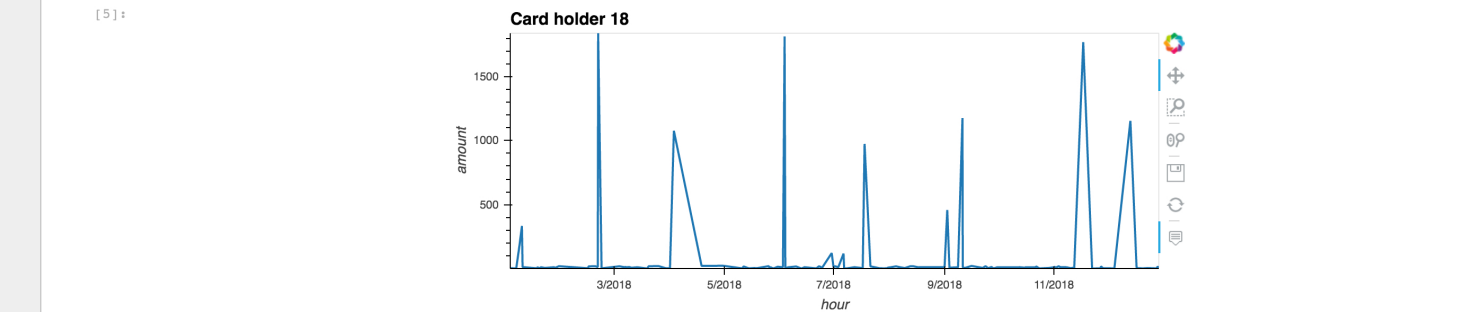

```
[6]:  # combined plot for card holders 2 and 18
      plot_1 = data_2_plot * data_18_plot
      plot_1.opts(
          title="Transactions Analysis for Card Holders 2 and 18",
          xlabel="Datetime",
          ylabel="Transaction Amount",
      )
```

[6]:
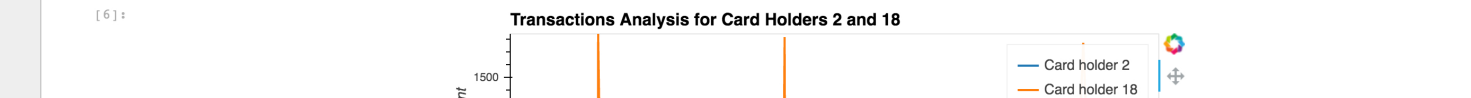
Sample Conclusions for Question 1

After visualy analyzed the plots and the spending patterns, it can be concluded that there may be some fraudulent transactions in the card holder 18 records since there are some anomalous amounts throughout the year that break the typical spending pattern that can be seen on card holder 2.

Data Analysis Question 2

Use `Plotly Express` to create a series of six box plots, one for each month, in order to identify how many outliers could be per month for **card holder id 25**. By observing the consumption patters, do you see any anomalies? Write your own conclusions about your insights.

```python
[7]: # loading data of daily transactions from jan to jun 2018 for card holder 25
query = """
        SELECT date_part('month', t.date) AS month, date_part('day', t.date) as day, t.amount
        FROM transaction AS t
        JOIN credit_card AS cc ON cc.card = t.card
        JOIN card_holder AS ch ON ch.id = cc.id_card_holder
        WHERE ch.id = 25 AND date_part('month', t.date) <= 6
        ORDER BY month, day
        """
df_question2 = pd.read_sql(query, engine)
```

```python
[8]: # loop to change the numeric month to month names
for i in range(df_question2.shape[0]):
    df_question2.iloc[i, 0] = calendar.month_name[int(df_question2.iloc[i, 0])]

df_question2.head()
```
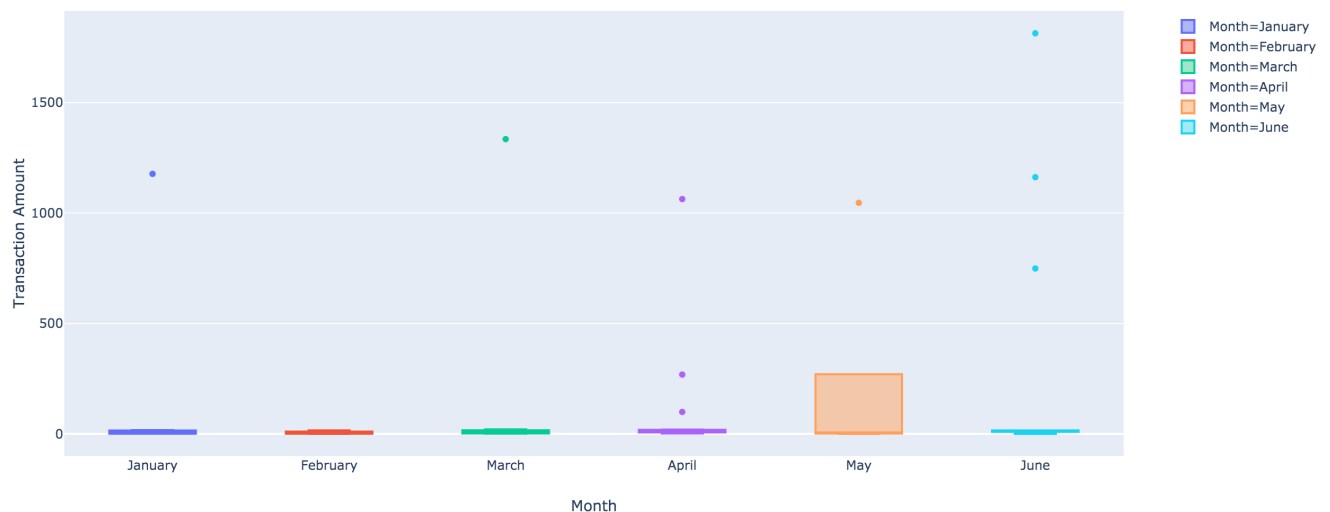
[8]:

	month	day	amount
0	January	2.0	1.46
1	January	5.0	10.74
2	January	7.0	2.93
3	January	10.0	1.39
4	January	14.0	17.84

```python
[9]: # creating the six box plots using plotly express
px.box(
    df_question2,
    x="month",
    y="amount",
    title="Monthly Transactions Analysis for Card Holder 25",
    labels={"month": "Month", "amount": "Transaction Amount"},
    color="month",
    boxmode="overlay",
)
```


Sample Conclusions for Question 2

It can be concluded that card holder 25 has been hacked along all the first semester of 2018, exept for february where there are not anomalous transactions.