# visual_data_analysis

May 4, 2021

## 1 Visual Data Analysis of Fraudulent Transactions

Your CFO has also requested detailed trends data on specific card holders. Use the starter notebook to query your database and generate visualizations that supply the requested information as follows, then add your visualizations and observations to your markdown report.

[1]:
```
# Initial imports
import pandas as pd
import calendar
import plotly.express as px
import hvplot.pandas
from datetime import datetime
from sqlalchemy import create_engine
from dotenv import load_dotenv
import os
```

[2]:
```
# Create a connection to the database
load_dotenv('../../.env')
db_url = os.getenv("db_url")
engine = create_engine(db_url)
```

### 1.1 Data Analysis Question 1

The two most important customers of the firm may have been hacked. Verify if there are any fraudulent transactions in their history. For privacy reasons, you only know that their cardholder IDs are 2 and 18.

- Using hvPlot, create a line plot representing the time series of transactions over the course of the year for each cardholder separately.
- Next, to better compare their patterns, create a single line plot that containins both card holders' trend data.
- What difference do you observe between the consumption patterns? Does the difference suggest a fraudulent transaction? Explain your rationale in the markdown report.

[3]:
```
# loading data for card holder 2 and 18 from the database
# Write the query
query = """
        select t.*, c.card_holder_id
        from transaction t
        join credit_card c on t.card = c.card
        where c.card_holder_id in (2, 18, 25)
        """
# Create a DataFrame from the query result. HINT: Use pd.read_sql(query, engine)
df = pd.read_sql(query, engine)
df.head()
```

```
[3]:      id                date  amount          card  id_merchant  card_holder_id
     0   567 2018-01-01 23:15:10    2.95  4.498003e+12           64              18
     1  2083 2018-01-02 02:06:21    1.46  4.319654e+12           93              25
     2  1552 2018-01-05 06:26:45   10.74  3.724148e+14           86              25
     3  2077 2018-01-05 07:19:27    1.36  3.441196e+14           30              18
     4  2439 2018-01-06 02:16:41    1.33  4.866761e+18          127               2
```

```
[4]: # Plot for cardholder 2
     card_holder_2 = df.loc[df['card_holder_id'] == 2].sort_values(by = 'date')
     plot_trans_2 = card_holder_2.hvplot.line(x='date', y='amount', rot=90,␣
      ↪color='green',
                                               title='Transactions 2018 for␣
      ↪Cardholder 2', xlabel='Date and Time', ylabel='Amount ($)',
                                          line_color='green')
     plot_trans_2
```

```
[4]: :Curve   [date]   (amount)
```

```
[5]: # Plot for cardholder 18
     card_holder_18 = df.loc[df['card_holder_id'] == 18].sort_values(by = 'date')
     plot_trans_18 = card_holder_18.hvplot.line(x='date', y='amount', rot=90,␣
      ↪color='green',
                                               title='Transactions 2018 for␣
      ↪Cardholder 18', xlabel='Date and Time', ylabel='Amount ($)',
                                            line_color='red')
     plot_trans_18
```

```
[5]: :Curve   [date]   (amount)
```

```
[6]: # Combined plot for card holders 2 and 18
     plot_trans_2 * plot_trans_18.opts(title='Transactions Overlay 2018 for␣
      ↪Cardholder 2 and 18')
```

```
[6]: :Overlay
        .Curve.I  :Curve   [date]   (amount)
        .Curve.II :Curve   [date]   (amount)
```

Conclusions: Credit card holder 2 is using super small amounts (<$19.51). Therefore, card holder 18 is riskier than the credit card holder 2 (~$1,000.00)

## 1.2 Data Analysis Question 2

The CEO of the biggest customer of the firm suspects that someone has used her corporate credit card without authorization in the first quarter of 2018 to pay quite expensive restaurant bills. Again, for privacy reasons, you know only that the cardholder ID in question is 25.

- Using Plotly Express, create a box plot, representing the expenditure data from January 2018 to June 2018 for cardholder ID 25.
- Are there any outliers for cardholder ID 25? How many outliers are there per month?
- Do you notice any anomalies? Describe your observations and conclusions in your markdown report.

```
[7]: # loading data of daily transactions from jan to jun 2018 for card holder 25
     # Write the query
     # Create a DataFrame from the query result. HINT: Use pd.read_sql(query, engine)
     df['date'] = pd.to_datetime(df['date']).dt.date
     card_holder_25 = df.loc[(df['card_holder_id']==25) & (df['date'] < datetime.
      ↪date(datetime(2018,7,1)))].sort_values(by = 'date')
     card_holder_25.head()
```

```
[7]:       id        date  amount          card  id_merchant  card_holder_id
     1   2083  2018-01-02    1.46  4.319654e+12           93              25
     2   1552  2018-01-05   10.74  3.724148e+14           86              25
     7   2108  2018-01-07    2.93  4.319654e+12          137              25
     11   754  2018-01-10    1.39  3.724148e+14           50              25
     13  3023  2018-01-14   17.84  3.724148e+14           52              25
```

```
[8]: # loop to change the numeric month to month names
     card_holder_25['monthnumber'] = pd.DatetimeIndex(card_holder_25['date']).month
     card_holder_25['monthName'] = card_holder_25['monthnumber'].apply(lambda x:␣
      ↪calendar.month_name[x])
     card_holder_25.drop(columns = ['date', 'monthnumber','id'], inplace = True)
     card_holder_25.head()
```

```
[8]:     amount          card  id_merchant  card_holder_id monthName
     1     1.46  4.319654e+12           93              25   January
     2    10.74  3.724148e+14           86              25   January
     7     2.93  4.319654e+12          137              25   January
     11    1.39  3.724148e+14           50              25   January
     13   17.84  3.724148e+14           52              25   January
```

```
[9]: # Creating the six box plots using plotly express
     px.box(card_holder_25, x = 'monthName', y = 'amount', notched = True)
```

Card holder 25 is quite suspicious. Every month there are some small amounts, spending over $1000.