Retail Sales Data Analysis

```
#run this code first to connect to the database and verify the connection is working
## DO NOT MODIFY THIS CODE BLOCK
## If you have placed this notebook in the jupyter notebooks folder properly,
## this block should return the first two rows of the customers table

from sqlalchemy import create_engine
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from IPython.core.interactiveshell import InteractiveShell
InteractiveShell.ast_node_interactivity = "all"

%matplotlib inline

cnxn_string = ("postgresql+psycopg2://{username}:{pswd}"
               "@{host}:{port}/{database}")
print(cnxn_string)

engine = create_engine(cnxn_string.format(
    username="postgres",
    pswd="behappy",
    host="postgres",
    port=5432,
    database="sqlda"))

engine.execute("SELECT * FROM customers LIMIT 2;").fetchall()
```

```
postgresql+psycopg2://{username}:{pswd}@{host}:{port}/{database}
[(1, None, 'Arlena', 'Riveles', None, 'ariveles0@stumbleupon.com', 'F', '98.36.172.246', None, None, None, None, None, None, None,
datetime.datetime(2017, 4, 23, 0, 0)),
 (2, 'Dr', 'Ode', 'Stovin', None, 'ostovin1@npr.org', 'M', '16.97.59.186', '314-534-4361', '2573 Fordem Parkway', 'Saint Louis', 'MO',
'63116', 38.5814, -90.2625, datetime.datetime(2014, 10, 2, 0, 0))]
```

Scenario

You are a team of extremely successful data scientists at a top motor dealership company. You need to create summary tables and visualizations that your boss will present at the next company shareholder meeting. She has sent you the following e-mail describing what she needs.

From: importantboss@topmotordealershipcompany.com

To: datascienceteam@topmotordealershipcompany.com

Subject: Data request for shareholder meeting

For our next shareholder meeting, we need to provide more information about sales performance across states, across dealerships, and across sales channels. Please send me information to address the following items for our next shareholder meeting along with your thoughts.

1. Sales performance at the state level (top 5 and bottom 5 states)
2. For the best performing states, which dealerships are performing well and how are they trending?
3. In states with dealerships, how has the distribution of sales amounts changed over time for different channels (internet vs. dealership) and sales types (low, typical, high value)?

Thank you!

-Important Boss

Your team promptly comes up with the following plan.

Part 1: Visualizing the top and bottom performing states

1. Write a SELECT query that returns the total sales amount for each state from January 1, 2016 to now. The table should have two columns, `state` and `total_sales_amount`, with one row for each state ordered by `total_sales_amount` in *descending* order. Make sure that

`total_sales_amount` is rounded appropriately. Attribute sales to states based on the **state in which the customer that made the purchase resides**. This way we can capture both sales made through dealerships, as well as sales made through our website, in evaluating state-level performance.

2. Use SQLAlchemy to execute the query and store the results in a pandas dataframe called `sales_by_state`.
3. Display the rows in `sales_by_state` corresponding to the 5 states with the **largest** total sales amount in *descending* order.
4. Display the rows in `sales_by_state` corresponding to the 5 states with the **smallest** total sales amount in *ascending* order.
5. Visualize sales performance by state for the top and bottom performing states discovered in 1.3 and 1.4. You can use more than one visualization. These should be **presentation ready** (e.g. appropriate and complete titles and axis labels, remove unnecessary/distracting features, display date range for total sales, no overlapping axis labels, etc.).

Include the code needed for each component of part 1 in the appropriate code block below.

```
#1.1
query = """SELECT c.state, ROUND(SUM(s.sales_amount))AS total_sales_amount
FROM customers AS c
INNER JOIN sales AS s
ON c.customer_id = s.customer_id
WHERE date_added::DATE >= '2016-01-01' AND state is not null
GROUP BY state
ORDER BY total_sales_amount DESC"""


#1.2 create dataframe
sales_by_state = pd.read_sql_query(query,engine)


#1.3 display top 5 performing states
sales_by_state[:5]
```

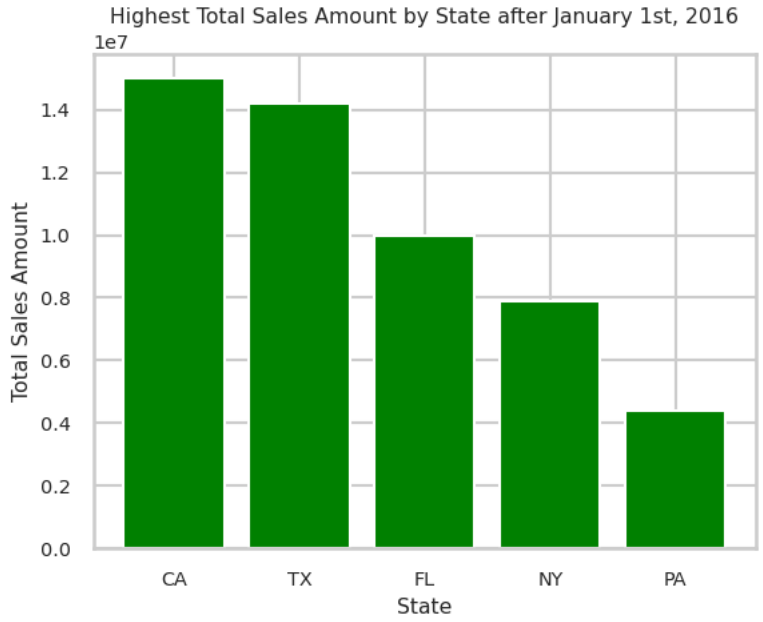
	state	total_sales_amount
0	CA	15005479.0
1	TX	14194945.0
2	FL	9990479.0
3	NY	7870301.0
4	PA	4387651.0

```
#1.4 display bottom 5 performing states
sorted_sales_by_state=sales_by_state.sort_values(by='total_sales_amount',ascending=True)
sorted_sales_by_state[:5]
```

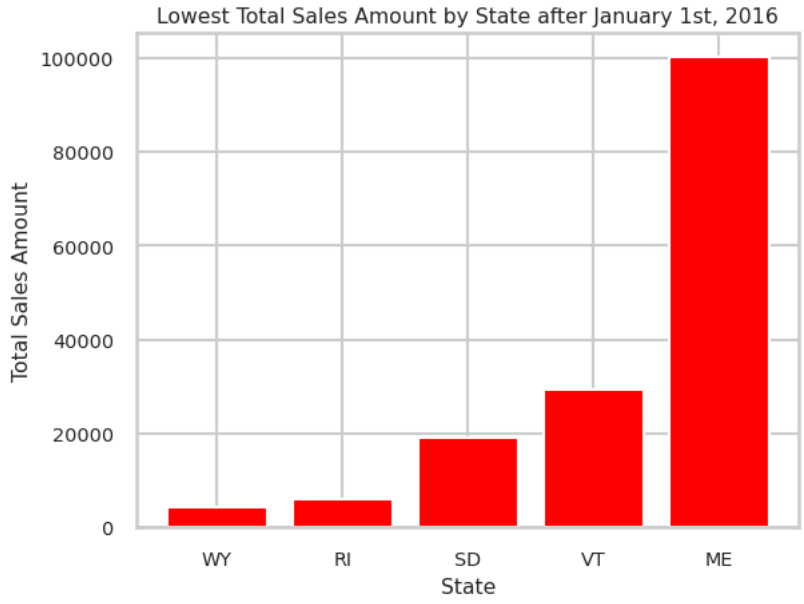
	state	total_sales_amount
50	WY	4200.0
49	RI	5980.0
48	SD	19060.0
47	VT	29400.0
46	ME	100195.0

```
#1.5 visualize top and bottom performing states
#Top performing states
state_top = ['CA', 'TX', 'FL', 'NY', 'PA']
total_sales_amount_top = [15005479, 14194945, 9990479, 7870301, 4387651]
plt.title('Highest Total Sales Amount by State after January 1st, 2016')
plt.xlabel('State')
plt.ylabel('Total Sales Amount')
plt.bar(state_top, total_sales_amount_top, color=['green', 'green','green', 'green','green'])
plt.show()

#Bottom performing states
state_bottom = ['WY','RI','SD','VT','ME']
total_sales_amount_bottom = [4200, 5980, 19060, 29400, 100195]
plt.title('Lowest Total Sales Amount by State after January 1st, 2016')
plt.xlabel('State')
plt.ylabel('Total Sales Amount')
plt.bar(state_bottom, total_sales_amount_bottom, color=['red', 'red', 'red', 'red', 'red'])
plt.show()
```

```
Text(0.5, 1.0, 'Highest Total Sales Amount by State after January 1st, 2016')
Text(0.5, 0, 'State')
Text(0, 0.5, 'Total Sales Amount')
<BarContainer object of 5 artists>
```




```
Text(0.5, 1.0, 'Lowest Total Sales Amount by State after January 1st, 2016')
Text(0.5, 0, 'State')
Text(0, 0.5, 'Total Sales Amount')
<BarContainer object of 5 artists>
```



Part 2: Top performing dealerships

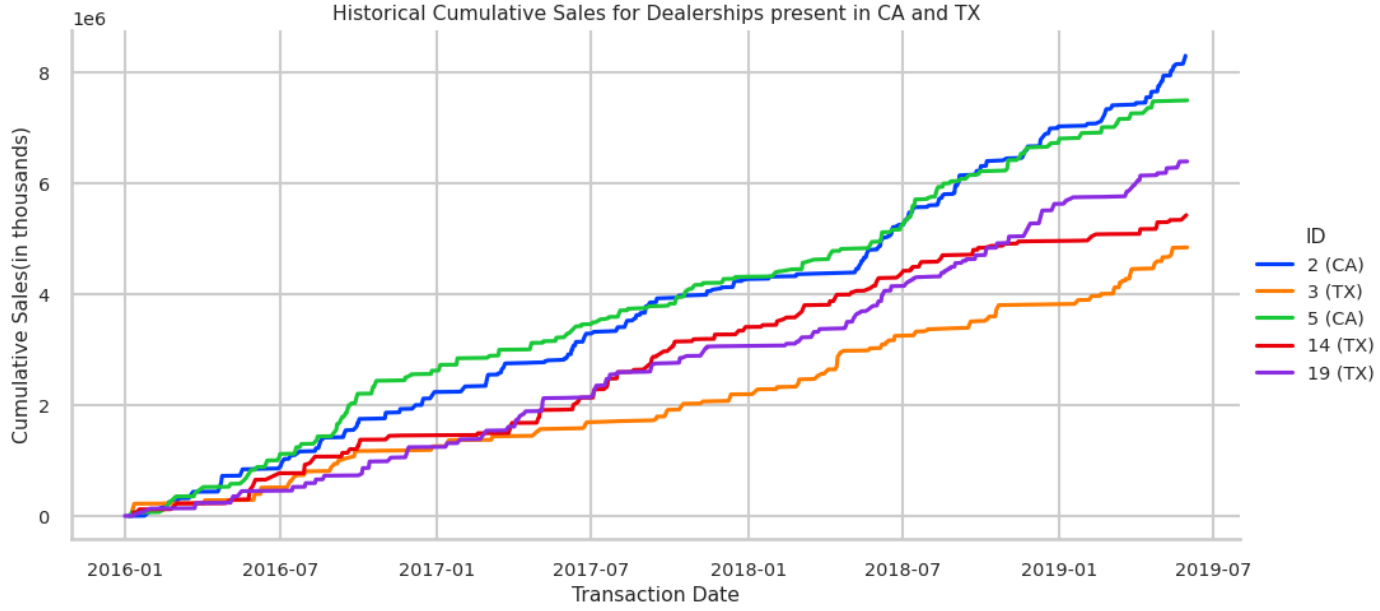
Create a table and visualization of historical cumulative sales amounts by dealership from January 1, 2016 to now. Only include dealerships located in the *top two* states determined in Part 1. It is OK to reference these two states by their abbreviations (e.g. AL, MS, WY) in the query you will develop below since this is a one-off request.

To do this, perform the following steps:

1. Write a SELECT query that returns three columns: `dealership_id`, `state`, `sales_transaction_date`, and `cumulative_sales`. `cumulative_sales` represents the cumulative sales amount from January 1, 2016 to the `sales_transaction_date` for dealership identified by `dealership_id`. There should be a row for each distinct combination of `dealership_id` and `sales_transaction_date` in the `sales` table (*hint*: window function).
2. Use SQLAlchemy to execute the query and store the results in a pandas dataframe called `cumulative_sales_bydealership`.
3. Appropriately visualize historical cumulative sales by dealership across sales transaction dates *in a single plot* (*hint*: seaborn). Visualization should be **presentation ready** (e.g. appropriate and complete titles and legend/axis labels, remove unnecessary/distracting features, display date range for total sales, no overlapping axis labels, integer-valued dealership IDs, states indicated clearly, variable names like `dealership_id` and `state` are replaced with appropriate text like 'ID' and 'State', etc.).

```
query = """
SELECT d.dealership_id as dealership_id,date(s.sales_transaction_date) as sales_transaction_date, d.state ,
SUM(sales_amount) OVER (PARTITION BY d.dealership_id ORDER BY sales_transaction_date) AS cumulative_sales
FROM dealerships as d
LEFT JOIN sales as s
ON d.dealership_id = s.dealership_id
WHERE sales_transaction_date >= '2016-01-01 00:00:00'
AND state in ('CA', 'TX')
ORDER BY dealership_id, sales_transaction_date;
"""


#2.2
cumulative_sales_bydealership = pd.read_sql_query(query,engine)


cumulative_sales_bydealership['variable'] = cumulative_sales_bydealership['dealership_id'].astype(str) + ' (' + cumulative_sales_bydealershi
plot=sns.relplot(x ="sales_transaction_date", y ="cumulative_sales",
            data= cumulative_sales_bydealership, kind= "line", hue="variable",aspect=2, palette="bright")
plt.title('Historical Cumulative Sales for Dealerships present in CA and TX')
plt.xlabel('Transaction Date')
plt.ylabel('Cumulative Sales(in thousands)')
plot._legend.set_title('ID')
plt.show()
```

```
Text(0.5, 1.0, 'Historical Cumulative Sales for Dealerships present in CA and TX')
Text(0.5, 31.87500000000003, 'Transaction Date')
Text(38.56335156249999, 0.5, 'Cumulative Sales(in thousands)')
```



Part 3: Sales amount by sales channel and sales type

Create tables and visualizations to compare sales amounts by sales channel for sales made on or after January 1, 2016 and before January 1, 2019. **Only include sales made to customers that reside in a state that has a dealership.** To do this, perform the following steps:

1. Write a SELECT query that returns sales with a transaction date on or after January 1, 2016 and before January 1, 2019 from the `sales` table made to customers that reside in a state that has a dealership. This table should have the following four columns: `channel`, `sales_amount`, and `sales_type` and `sales_year`. `channel` and `sales_amount` are exactly as appears in the `sales` table. `sales_type` is a derived categorical field that takes on a value of `'High value'` when `sales_amount` is above 50000, `'Typical value'` when `sales_amount` is above 10000 but less than or equal to 50000, and `'Low value'` when `sales_amount` is less than 10000. `sales_year` is the year from the `sales_transaction_date` field.

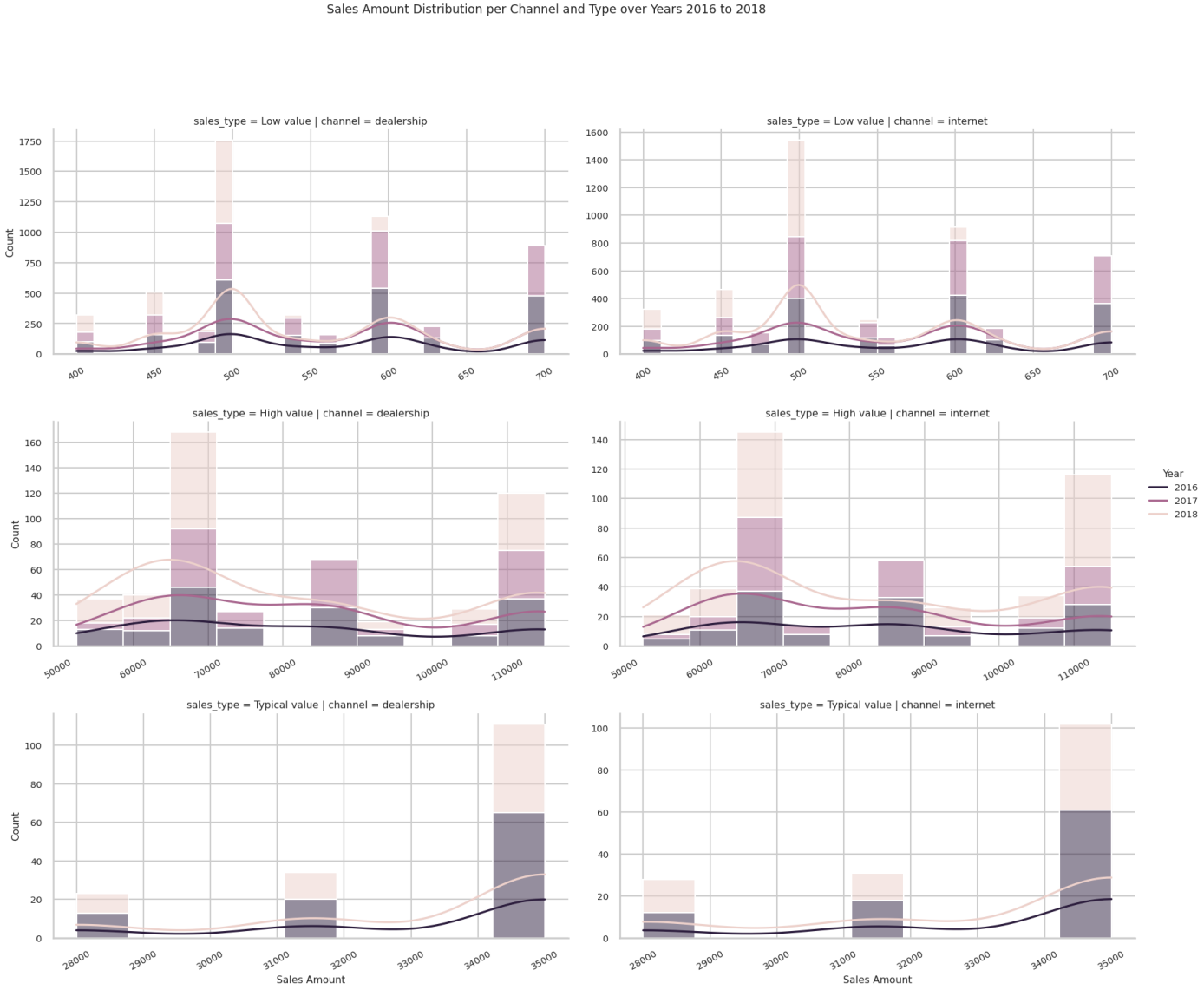
2. Use SQLAlchemy to execute the query and store the results in a pandas dataframe called `sales_from_dealershipstates`.

3. Appropriately visualize the distribution of sales amounts and how it changes by `channel`, `sales_year`, and `sales_type`. To do this, create multiple plots, one for each distinct combination of `channel` and `sales_type`. For each plot, visualize and compare the distribution of sales amounts for each sales year (2016, 2017, 2018) by superimposing these yearly distributions on the same plot. For example, one plot will visualize distribution of sales amounts in 2016, 2017, and 2018 for low value internet sales. Arrange the plots so that you can see changes across `channel` and `sales_type` (*Hint: seaborn.FacetGrid*). Visualizations should be **presentation ready** (e.g. appropriate and complete titles and legend/axis labels, remove unnecessary/distracting features, display date range for total sales, no overlapping axis labels, replace variable names like `sales_year` with appropriate text like 'Year', etc.).

```
#3.1 select query
query = """
select s.channel, s.sales_amount,
case
when s.sales_amount > 50000 then 'High value'
when s.sales_amount > 10000 and s.sales_amount <=50000 THEN 'Typical value'
else 'Low value'
end as sales_type,
extract(year from s.sales_transaction_date) as sales_year
from sales s
join customers c on s.customer_id = c.customer_id
where
s.sales_transaction_date >= '2016-01-01' and
s.sales_transaction_date < '2019-01-01' and
c.state in (select distinct state from dealerships)
"""
```

```
#3.2 create data frame
sales_from_dealershipstates = pd.read_sql(query, engine)


#3.3 visualization
# My custom color palette
custom_palette = sns.color_palette("husl", 3)

sns.set_theme(context='talk', style='whitegrid', palette=custom_palette, font='sans-serif', font_scale=.625, color_codes=True, rc=None)

g = sns.FacetGrid(sales_from_dealershipstates, col="channel", row="sales_type",
height=5.5, aspect=1.8, sharex=False, sharey=False, palette=custom_palette,
gridspec_kws={"hspace": 0.3, "wspace": 0.1})

def new_displot(x, hue, **kwargs):
    sns.histplot(x=x, hue=hue, **kwargs, kde=True, multiple="stack")


g.map_dataframe(new_displot, 'sales_amount', 'sales_year', hue_order=['2016', '2017', '2018'])
g.fig.suptitle("Sales Amount Distribution per Channel and Type over Years 2016 to 2018", y=1)
g.set_axis_labels("Sales Amount", "Count")

for ax in g.axes.flat:
    ax.tick_params(axis='x', which='both', labelbottom=True, rotation=30)

legend_labels = {year: color for year, color in zip(['2016', '2017', '2018'], custom_palette)}
g.add_legend(title="Year", label_order=['2016', '2017', '2018'], labels=legend_labels)

plt.show()
```

```
/usr/local/lib/python3.10/site-packages/seaborn/axisgrid.py:118: UserWarning: This figure includes Axes that are not compatible with tig
  self._figure.tight_layout(*args, **kwargs)
/usr/local/lib/python3.10/site-packages/seaborn/axisgrid.py:118: UserWarning: This figure includes Axes that are not compatible with tig
  self._figure.tight_layout(*args, **kwargs)
<seaborn.axisgrid.FacetGrid at 0x7fa2f58c09d0>
Text(0.5, 1, 'Sales Amount Distribution per Channel and Type over Years 2016 to 2018')
<seaborn.axisgrid.FacetGrid at 0x7fa2f58c09d0>
/usr/local/lib/python3.10/site-packages/seaborn/axisgrid.py:181: UserWarning: You have mixed positional and keyword arguments, some inpu
  figlegend = self._figure.legend(handles, labels, **kwargs)
<seaborn.axisgrid.FacetGrid at 0x7fa2f58c09d0>
```



Part 4: Takeaways from the analysis

Provide your thoughts about the analysis above by answering the following questions in the blank markdown cells provided below. No code should be run for this section.

1. (Part 1) What are some potential hypotheses as to why the top 5 performing states have the highest sales amounts? Describe how you would test your hypotheses in further analysis. Do not conduct any additional analyses or write any more queries, just describe in words.

2. (Part 1) What are some potential hypotheses as to why the bottom 5 performing states have the lowest sales amounts? Describe how you would test your hypotheses in further analysis. Do not conduct any additional analyses or write any more queries, just describe in words.

3. (Part 2): How would you characterize the historical performance of the dealerships visualized in Part 2 (e.g. good, bad, growing, declining, etc.)? Describe some of the trends in relative performance over time for the dealerships. Be specific and cite specific elements of the visualization created in Part 2 to support your claims. Specify any additional factors you would want to consider that would influence your performance assessment.

4. (Part 3): How does the distribution of sales amounts change from one year to the next for each channel and sale type? Are the number of transactions for certain channels and sales types increasing or decreasing over time? For each sale type and channel, is the distribution of sales amounts changing over time (e.g. Are the typical sales amounts for low value internet sales shifting from 2016 to 2018? If so, how are the values shifting over time?) Be specific and cite specific elements of the visualization created in Part 3 to support your claims.

5. (Part 3): What are some potential hypotheses as to why the distribution of sales amounts compared across channel, year, and sales type behaves in the manner you described in 4.4? Describe how you would test your hypotheses in further analysis. Do not conduct any additional analyses or write any more queries, just describe in words.

Part 4 Responses

For each of the following questions, answer in as much preciseness and clarity that you can. Refer back to the tables and plots that you have created to back up your answers if necessary. Answer each question in the cell below. You are NOT to code anything for this section. This is for you to reflect on the analysis developed in response to Parts 1-3.

1. (Part 1) What are some potential hypotheses as to why the top 5 performing states have the highest sales amounts? Describe how you would test your hypotheses in further analysis. Do not conduct any additional analyses or write any more queries, just describe in words.

One potential hypothesis includes that the top 5 performing states have a higher total population. This would result in higher sales in these states as there is a larger population making purchases. Another hypothesis could be that it is dependent on gender. For example, if there are more females in the state, this could result in more purchases. To test these hypotheses, we could compare the total population value to the number of customers in the state. If there are more people in California contributing to sales, we can attribute this to a higher population. Furthermore, we can group by gender to determine which group contributes to sales more.

2. (Part 1) What are some potential hypotheses as to why the bottom 5 performing states have the lowest sales amounts? Describe how you would test your hypotheses in further analysis. Do not conduct any additional analyses or write any more queries, just describe in words.

The main hypothesis for the lowest performing states would be these states have less total population. Another hypothesis could be is if there are less dealership sales as these states are not necessarily known for shopping, tourism, etc. To test this, it would be beneficial to count how many dealerships sales contributed versus how many did not.

3. (Part 2): How would you characterize the historical performance of the dealerships visualized in Part 2 (e.g. good, bad, growing, declining, etc.)? Describe some of the trends in relative performance over time for the dealerships. Be specific and cite specific elements of the visualization created in Part 2 to support your claims. Specify any additional factors you would want to consider that could influence your performance assessment.

From the visualisation we can see that dealerships 2, 3, 5, 14 and 19 are present in the state CA and TX, which are top performing states. Mostly the plot represents a growing trend with time this indicates that sales has increased gradually from the year 2016 and increased. When we can observe the plot overall we can observe that the dealerships 2 and 5 have a similar pattern, which suggests that sales is similar in that particular state which is CA. Dealerships 13, 14 and 19 have a similar pattern therefore it also suggests that the sales is similar for all dealerships in TX. But when look at each dealership closely we observe that, Dealership 14 experiences a decline after March 2018. Dealership experiences a sudden. growth after Jan 2017. Dealership 19 has a spiky trend as we can observe a lot of fluctuations. Dealership 2 follows a

similar fluctuating pattern. Dealership 3 is only dealership of the five dealerships which follows a steady growth from 2016 to 2019 without much of a decline. Factor to consider that will better describe the performance is the Sales Amount because sometimes the even the sales quantity might be less the profit earned can more. So if the Sales Amount was also use in the visualisation we might have observed a different pattern.

4. (Part 3): How does the distribution of sales amounts change from one year to the next for each channel and sale type? Are the number of transactions for certain channels and sales types increasing or decreasing over time? For each sale type and channel, is the distribution of sales amounts changing over time (e.g. Are the typical sales amounts for low value internet sales shifting from 2016 to 2018? If so, how are the values shifting over time?) Be specific and cite specific elements of the visualization created in Part 3 to support your claims.

For the low value sales, both dealerships and internet sales show an increase in sales amount between 2016 and 2017 but a decrease from 2017 and 2018. Trends in low-value dealership sales are pretty similar to those for low-value internet sales Overall sales count for low-value sales decreased from 2017 to 2018. The $400-500$ range had an increase in sales count for 2018, but again there is a decline in sales within the $500-700$ range. The distribution of sales count also increased a bit but remained relatively stable between 2016 and 2017.

For the high value sales, both dealerships and internet sales also show an increase in sales amount between 2016 and 2018. Overall sales count for high-value sales increased from 2016 to 2018. The $65000-70000$ range had an increase in sales count for all three years. The distribution of sales count also increased a bit but remained relatively stable between 2016 and 2017.

For the typical value sales, both dealerships and internet sales show an increase in sales amount due from 2016 to 2018. Overall sales count for high-value sales increased from 2016 to 2018. The $34000-35000$ range had an increase in sales count for all three years. The distribution of sales count also decreased a bit between 2016 and 2018.

For low value sales, both dealership and internet sales have decreased. As seen by the line graph, the count has decreased over time. For high value sales, both dealership and internet sales have maintained a steady level as the line graph is pretty stagnant. For typical values, both dealership and internet sales have increased as seen by the increasing line graph.

For low value sales, both dealership and internet sales have observed high sales at the average sales amounts of $500-600$.

For high value sales, both dealership and internet sales have observed high sales at the average sales amounts of $65000-70000$.

For typical value sales, both dealership and internet sales have observed high sales at the average sales amounts of $34000-35000$.

5. (Part 3): What are some potential hypotheses as to why the distribution of sales amounts compared across channel, year, and sales type behaves in the manner you described in 4.4? Describe how you would test your hypotheses in further analysis. Do not conduct any additional analyses or write any more queries, just describe in words.

Channel: The preference of the customer to buy through a channel might have changed over the years. People in the later years might prefer buying through internet as it is more accessible in this era where the internet is so easily available. This could be tested by conducting a survey of channel preference among common people and the reason for choosing it.