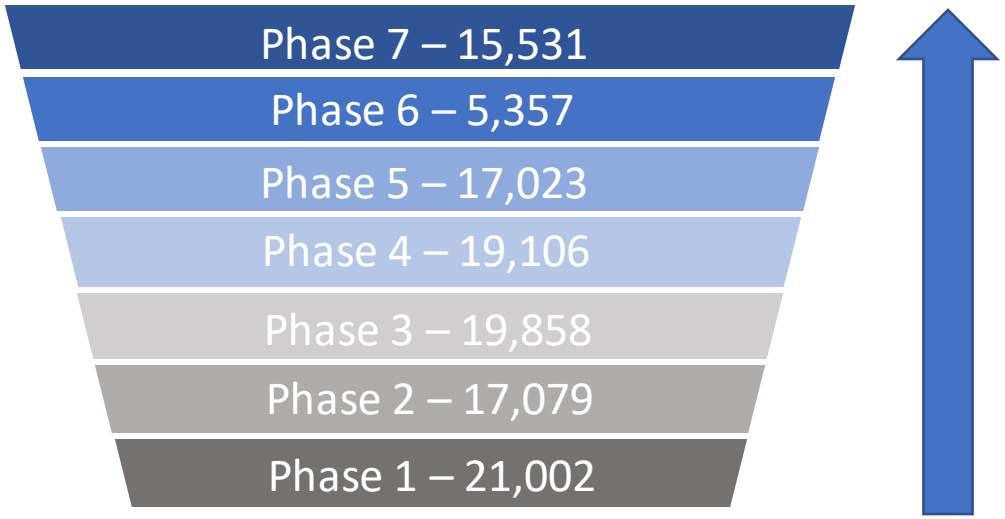
Harris County Case Study

Tai Tony Tran
Data Scientist
taitran2002@hotmail.com

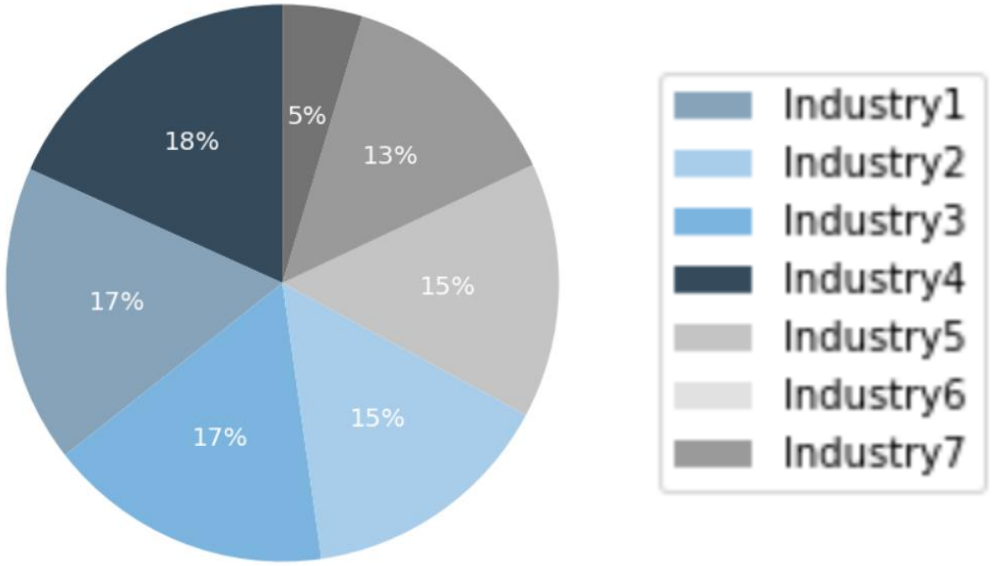
Client Sales Pipeline

Snapshot	Open	Closed Won	Close Lost	Total
Fiscal Year 2019 – 2020	90,310,121,319 Total Sales	10,060,167,028 Total Sales	2,531,162,528 Total Sales	110,465,430,331 Total Sales
Period 1 – 13	72,111 Count	25,947 Count	2,892 Count	107,536 Count

By Phase (Count)



By Industry



Which Products to Focus on Going Forward

Assumptions

- each product sale creates the same amount of profit in USD
- the company's metric for success is to maximize potential profit and sales

Key Takeaway

The client want to focus on products that **maximize** total **Weighted Sales** and show **positive** growth trends over time.

Total Weighted Sales

Top 5 Total Weight Sales by Item

Item Id	2019	2020	Total Weighted Sales
1650	526,978,559	1,157,333,961	1,684,312,520
2061	591,734,150	572,925,520	1,164,659,670
1818	375,658,065	730,103,540	1,105,761,605
2004	403,860,361	535,241,533	939,101,895
1629	357,677,003	448,149,619	805,826,622

1. Python Code in Appendix

These 37 Items have a Weighted Sales of 0

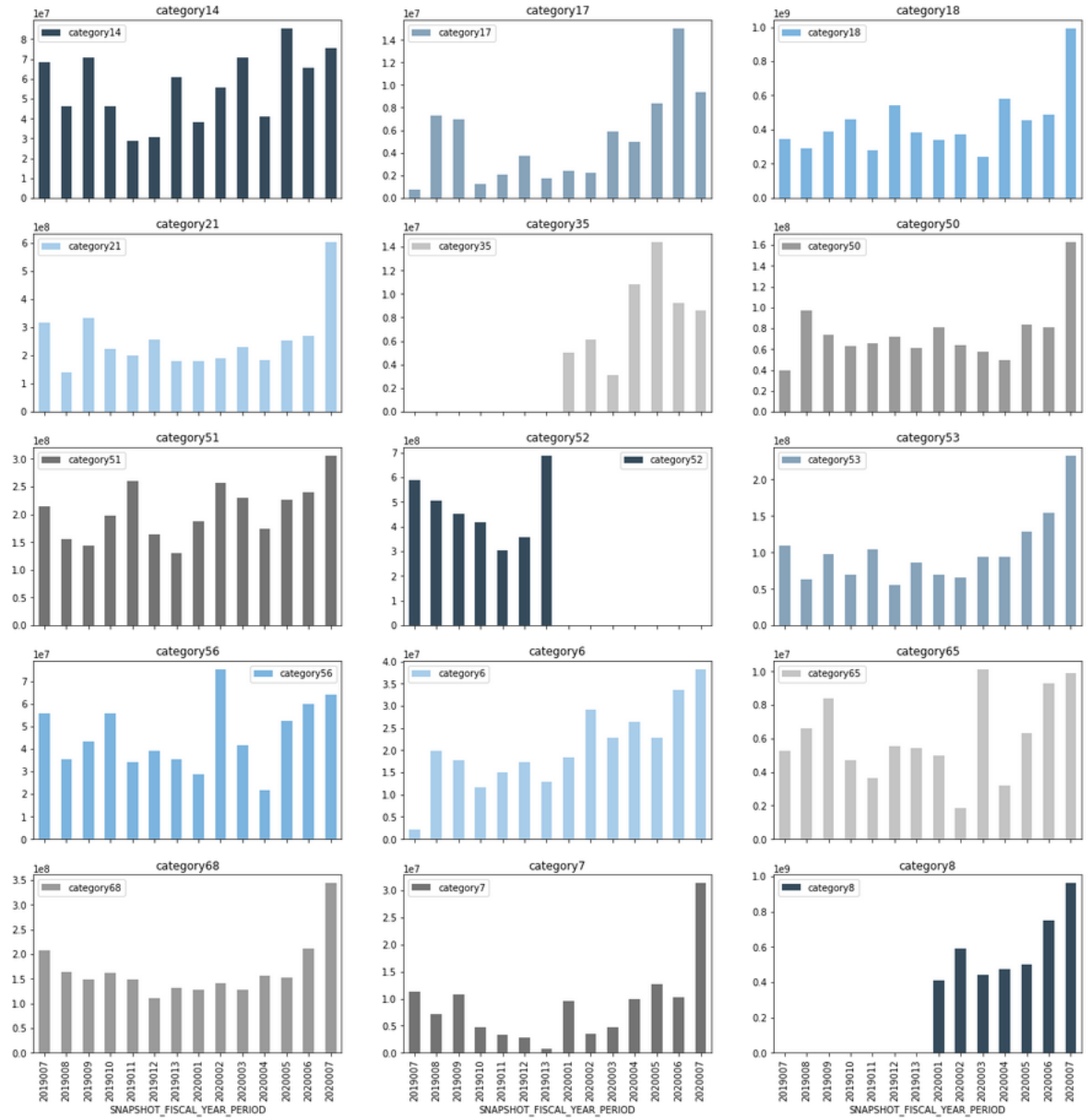
6015	21294	726	732	5943	5958
1266	801	1245	1233	15813	888
6051	6069	15438	1113	1092	1080
927	972	4647	4641	1680	657
25032	3198	3330	3387	24963	24918
4632	24768	3522	1839	3399	942
4629					

Highest Weighted Sales by Product_Category2_text

Item Id	Total Weighted Sales
category18	6,157,351,527
category8	4,125,766,049
category21	3,557,795,425
category52	3,307,325,336
category51	2,884,306,571

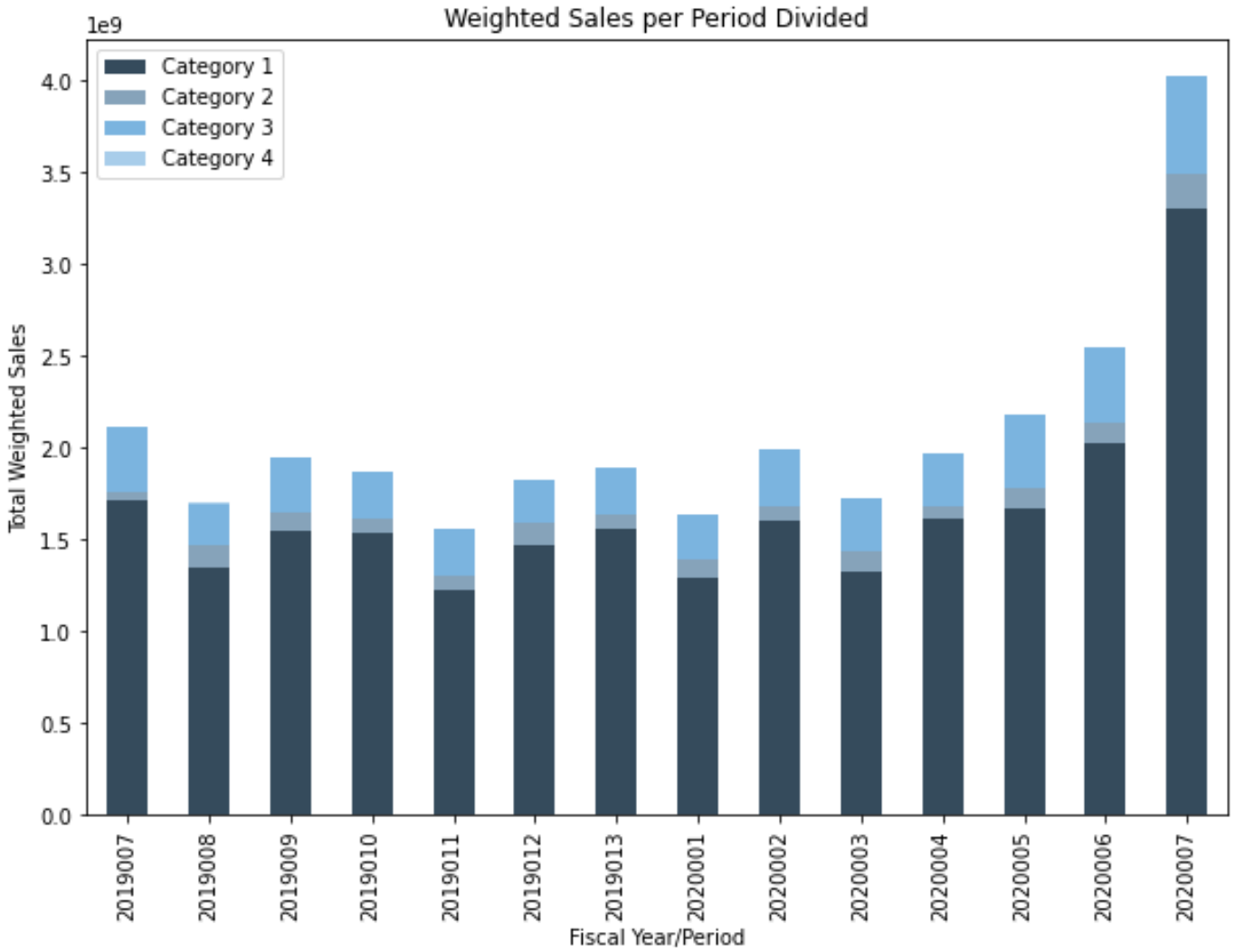
Categories 6, 7, 8, 14, 17, 18, 21, 35, 50, 51, 53, 56,65, 68 Show positive trends in this snapshot period.

Category 52 shows a negative trend despite the high total weighted sales.



2. Python Code in Appendix

Weighted Sales Separated by Product Mapping Category



Product Mapping Category	Average Weighted Sales	Count
Category 1	429,069.83	54,137
Category 2	51,683.86	26,534
Category 3	166,354.94	26,338
Category 4	18,125.63	527

Products Mapped into Category 1 had the highest Average Weighted Sales and highest count

3. Python Code in Appendix

Which Customers and Opportunities to Focus on Going Forward

- Client should focus on customers and opportunities that:
  1. Customers with a high count of opportunities that are "Closed" and "Won"
  2. Customers with a high total count of opportunities
  3. Customers with a large total weighted sales

What **region** are our 'best' customers from?

What **industry** are our 'best' customers from?

Should we focus on new or returning customers?

Top 10 Customers by Total Weighted Sale

Customer Id	Total Weighted Sales	Average Weighted Sales	Total # of Opportunities	Ratio of Won Opportunities
24018057	706,857,769	2,272,854	311	83 / 311 = 26.69%
24048879	598,351,669	600,152	997	224 / 997 = 22.47%
24035370	539,739,707	2,230,329	242	45 / 197 = 22.84%
24000954	455,511,484	1,980,484	230	35 / 230 = 15.22%
24043533	453,132,380	1,562,525	290	58 / 290 = 20.00%
24041244	440,527,906	1,034,103	426	33 / 426 = 7.75%
24036867	438,592,695	861,675	509	97 / 509 = 19.06%
24028941	420,689,030	600,127	701	76 / 701 = 10.84%
24043992	410,096,490	1,419,018	289	33 / 289 = 11.42%
24000882	394,500,355	1,860,850	212	35 / 212 = 16.51%

Customer Info

- Client should focus on customers from Industry 4 and US – East, due to a drastic difference in historic total weighted sales
- Both new customers and returning customers are have shown growth between 2019 and 2020.
- All Industries grew between 2019 and 2020.
- Only US – West decreased in total weighted sales between 2019-2020

New Clients

Client Type	Weighted Sales 2019	Weighted Sales 2020	% Change
New Customer	6,056,427,215	7,340,227,356	121.20%
Returning Customer	6,850,818,411	8,743,468,655	127.63%

Industry

Industry	Weighted Sales 2019	Weighted Sales 2020	% Change
Industry 1	1,433,774,372	2,184,317,648	152.35%
Industry 2	1,385,092,325	2,017,925,436	145.69%
Industry 3	1,154,204,501	1,336,945,558	115.83%
Industry 4	6,097,531,388	7,413,050,831	121.57%
Industry 5	1,613,699,515	1,662,796,085	103.04%
Industry 6	130,603,825	183,868,475	140.78%
Industry 7	1,092,339,696	1,284,791,974	117.62%

Region

Region	Weighted Sales 2019	Weighted Sales 2020	% Change
US - East	7,469,171,098	9,163,566,295	122.69%
US - Central	3,274,365,969	4,772,158,877	145.74%
US - West	1,966,580,583	1,863,201,604	94.74%
US – National Office	197,127,974	284,744,234	144.45%

4. Python Code in Appendix

Machine Learning Predictions

- Created a Random Forest Regressor model to predict **Weighted Sales** fit on columns such as:
  - Product category4 id
  - Total sales
  - Fiscal year and period
  - Customer industry
  - Opportunity id
  - Customer sector
  - Product Mapping Category
  - Region

Key takeaway

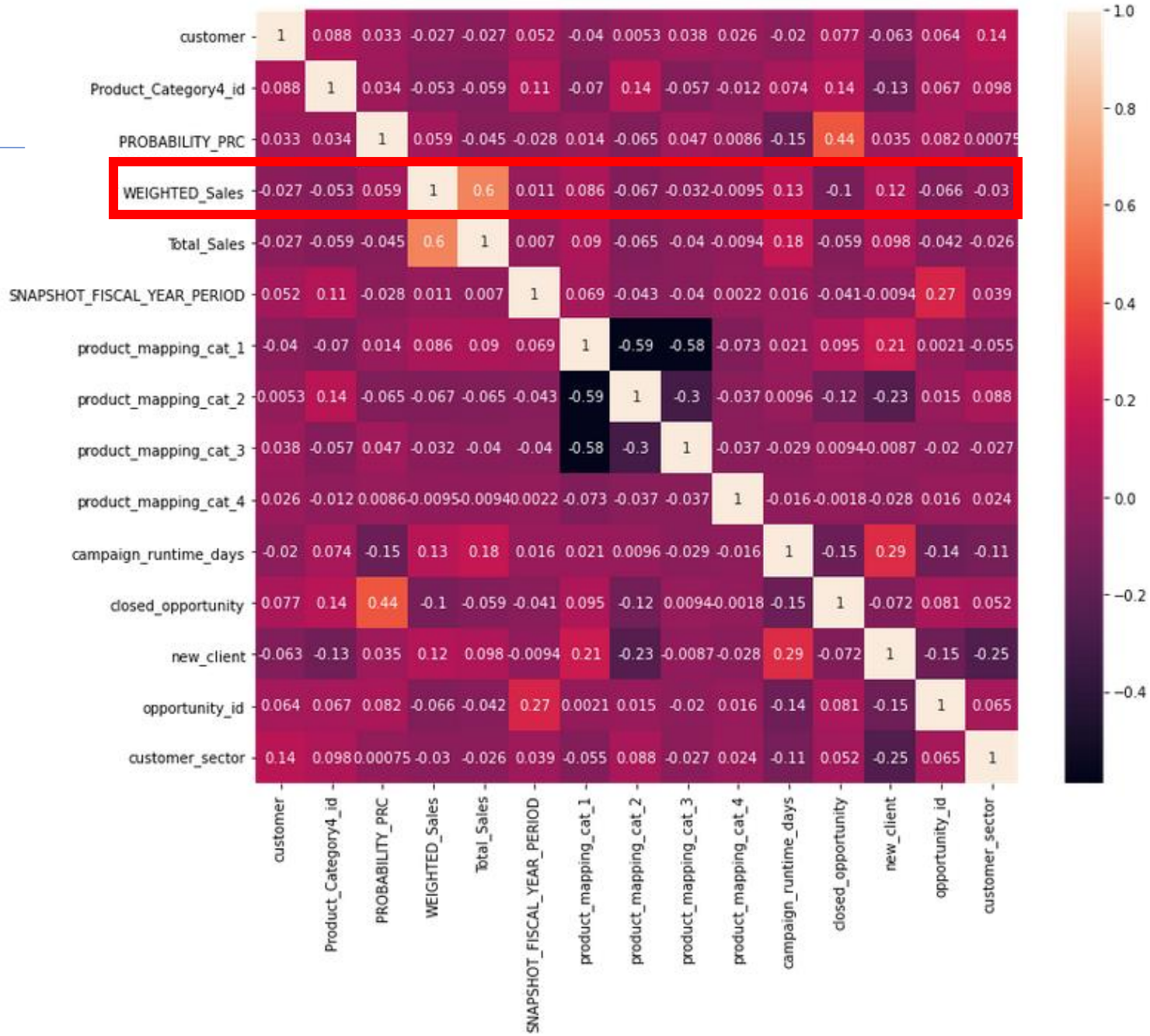
Metrics used to measure the accuracy of the model were mean absolute error and r^2.

An r^2 of 0.47 was recorded, meaning that 47% of the variance in the data are explained by the model created. A higher r^2 means a more accurate model, therefore a r^2 of 47% is good, however, it could be improved with more data and other features not included.

5. Python Code in Appendix

Correlation Heatmap

- There are no strong positive or negative correlation for Weighted Sales and other columns to use for machine learning predictions.



Where has the client seen the most success?

- The client has seen the most success with the growth of its new and recurring customer base in all recorded industries and most regions.
- The client has seen the largest percentage of growth with **recurring customers**, for customers in **Industry 1**, and customers in **US – East**.
- Client has seen the most success with Product_Category4_id of **1650**; not only does it have the highest total weighted sales, it has the highest increase in weighted sales over a 200% increase from 2019 to 2020

Appendix

1.	index = customer_model_df.groupby('Product_Category4_id')['WEIGHTED_Sales']\ .sum().sort_values(ascending=False).head().index customer_model_df[customer_model_df['Product_Category4_id'].isin(index)]\ .groupby(['fiscal_year', 'Product_Category4_id'])['WEIGHTED_Sales'].sum()
2.	customer_model_df.groupby(['SNAPSHOT_FISCAL_YEAR_PERIOD', 'Product_Category2_text'])\ ['WEIGHTED_Sales'].sum().sort_values(ascending=False).unstack()\ .plot(kind='bar', stacked=False, figsize=(15,100),\ subplots=True, layout=(24,2), color=colors) plt.title('Weighted Sales per Period Divided') plt.xlabel('Fiscal Year/Period') plt.ylabel('Total Weighted Sales') plt.show()

Appendix

3.	temp_df.groupby(['SNAPSHOT_FISCAL_YEAR_PERIOD', 'product_mapping_category'])['WEIGHTED_Sales']\ .sum().unstack().plot(kind='bar', stacked=True, figsize=(10,7),\ color=colors) plt.title('Weighted Sales per Product Mapping Category') plt.legend(['Category 1', 'Category 2', 'Category 3', 'Category 4']) plt.xlabel('Fiscal Year/Period') plt.ylabel('Total Weighted Sales') plt.show()
4.	customer_model_df.groupby(['fiscal_year', 'new_client'])['WEIGHTED_Sales'].sum() customer_model_df.groupby(['fiscal_year', 'customer_INDUSTRY_DESC'])['WEIGHTED_Sales'].sum() customer_model_df.groupby(['fiscal_year', 'customer_REGION_DESC'])['WEIGHTED_Sales'].sum()
5.	reg_rfor = make_pipeline(OneHotEncoder(handle_unknown='ignore'), SimpleImputer(strategy='most_frequent'), RandomForestRegressor(n_estimators=75, random_state=0, n_jobs=4)) reg_rfor.fit(X_train, y_train) y_pred3 = reg_rfor.predict(X_test) error3 = mean_absolute_error(y_test, y_pred3) r2_3 = r2_score(y_test, y_pred3) print(error3) print(r2_3)

Appendix

6.	plt.subplots(figsize=(12,10)) sns.heatmap(customer_model_df.corr(),annot=True) plt.show()
FULL JUPYTER NOTEBOOK	This project was coded in Python in a Jupyter Notebook. All my code can be found at the following github link: https://github.com/TonyTranPortfolio/Portfolio/blob/main/Harris%20County%20Case%20Study/Notebook/Harris%20County%20Data%20Analytics%20Case%20Study.ipynb