Melissa Young
DSC680 – T301 Applied Data Science
12.1 Project 3: White Paper

White Paper on Addressable Market Quantification and Sales Flywheel Analysis

Business Problem: The core aim of this project is to rigorously quantify and analyze the addressable market percentages across several key regions, beginning with a detailed focus on the Tampa Bay area in Florida. By investigating sales flywheel patterns, this project aims to derive actionable insights that will significantly enhance market penetration strategies, streamline sales processes, and predict flywheel locations.

Background/History: Understanding the addressable market and the dynamics of sales flywheels is crucial for companies aiming to optimize their market strategies. The concept of a sales flywheel revolves around creating a self-sustaining cycle where each phase of the sales process feeds into the next, building momentum and driving continuous growth. This approach contrasts with the traditional sales funnel, which often ends at the point of sale.

Data Explanation: The datasets were sourced from various systems including internal systems of record for the operations pertaining to MAV company sales and public data relating to family housing permitting and sales over the last 5 years.

1. **Warranty Sales List from MAV Corp**: Comprising detailed records of warranty sales linked to various builders and projects under MAV Corp, this dataset is pivotal for understanding warranty penetration rates and customer retention. https://github.com/Joven0218/DSC680---Applied-Data-Science/blob/main/Project%2010/hub_msa.xlsx (Tab = Warranty Sales 19)
2. **Tampa MSA Builders list from HubSpot**: This list includes comprehensive data on builders within the Tampa Metropolitan Statistical Area (MSA), detailing operational scale, specialty, and historical sales performance. https://github.com/Joven0218/DSC680---Applied-Data-Science/blob/main/Project%2010/hub_msa.xlsx (Tab = Tampa MSA Builders)
3. **Census Data for New Residential Sales**: Sourced from the U.S. Census Bureau (https://www.census.gov/construction/nrs/current/index.html), this dataset provides monthly updates on the volume of new residential sales, which is crucial for tracking market fluctuations and emerging trends.
4. **Census Data for New Residential Permitting**: Also from the U.S. Census Bureau (https://www.census.gov/construction/bps/current.html), this information reveals the rate of new residential permitting, serving as a leading indicator of construction and economic activity within the region.
5. **Region Mapping Definitions**: Essential for accurate geographical analysis, these definitions (https://www.census.gov/construction/soc/definitions.html) help standardize regional segmentation for mapping.

   See Appendix for data dictionaries.

Exploratory Data Analysis (EDA):

The project commenced with the loading and initial inspection of two distinct datasets: warranty_sales and tampa_msa_builders. The warranty_sales dataset contains detailed information about warranty sales linked to various builders and projects, whereas tampa_msa_builders includes data on builders within the Tampa Metropolitan Statistical Area (MSA).

1. Data Loading and Initial Exploration

Warranty Sales Data (warranty_sales):

- Loaded from an Excel file, containing fields like CertificateNumber, WarrantyEnrollmentAppID, ActivatedAt, and Premium.
- Initial exploration involved understanding data types, non-null counts, and basic statistics of numerical fields.

	ClosingDate	WarrantyEnrollmentDate	ActivatedAt	WarrantyLimit	BuilderNumber	BuilderLawsonCode	Premium
count	159797	159011	160463	1.601370e+05	1.604600e+05	160463.000000	160463.000000
mean	2021-06-08 07:32:08.649473792	2021-06-03 06:37:44.854632704	2021-07-04 12:24:17.363305216	3.906443e+05	1.293416e+07	6081.259131	494.543496
min	2001-05-25 04:00:00	2001-08-28 00:00:00	2019-01-04 00:00:00	0.000000e+00	1.190000e+02	6011.000000	0.000000
25%	2020-03-24 00:00:00	2020-03-20 00:00:00	2020-04-20 00:00:00	2.607850e+05	1.033000e+04	6077.000000	172.200000
50%	2021-05-21 00:00:00	2021-05-18 00:00:00	2021-06-17 00:00:00	3.359900e+05	1.550300e+04	6083.000000	348.840000
75%	2022-08-15 00:00:00	2022-07-29 00:00:00	2022-09-06 00:00:00	4.493750e+05	1.915200e+04	6083.000000	687.060000
max	2026-06-28 00:00:00	2026-06-28 00:00:00	2024-03-22 21:31:29.272000	5.000000e+06	2.300000e+11	6100.000000	15750.000000
std	NaN	NaN	NaN	2.286114e+05	1.722483e+09	5.544929	501.432144

Builder Data (tampa_msa_builders):

- Also loaded from an Excel file, featuring details such as BuilderName, City, State/Region, and Total Units.
- A similar exploratory data analysis was performed to assess data quality and structure.

2. Data Cleaning

Duplicate Removal:

- Datasets were checked for duplicate records using drop_duplicates(), ensuring that the analysis would be based on unique entries only.
- It was decided to drop null values from the dataset as it presented no loss of value.

Data Type Standardization:

- Ensured that state identifiers across datasets (StateId in warranty_sales and State/Region in tampa_msa_builders) were standardized by converting them to uppercase and stripping any extra whitespace.
- The cleaned dataset now had 158,010 rows and 16 columns.

3. Data Merging and Aggregation

State-Level Aggregation:

- Premiums and contract costs were aggregated by the state to analyze geographical distributions of market activity. This involved grouping by state and summing up premiums and contract costs, respectively.

Data Integration:

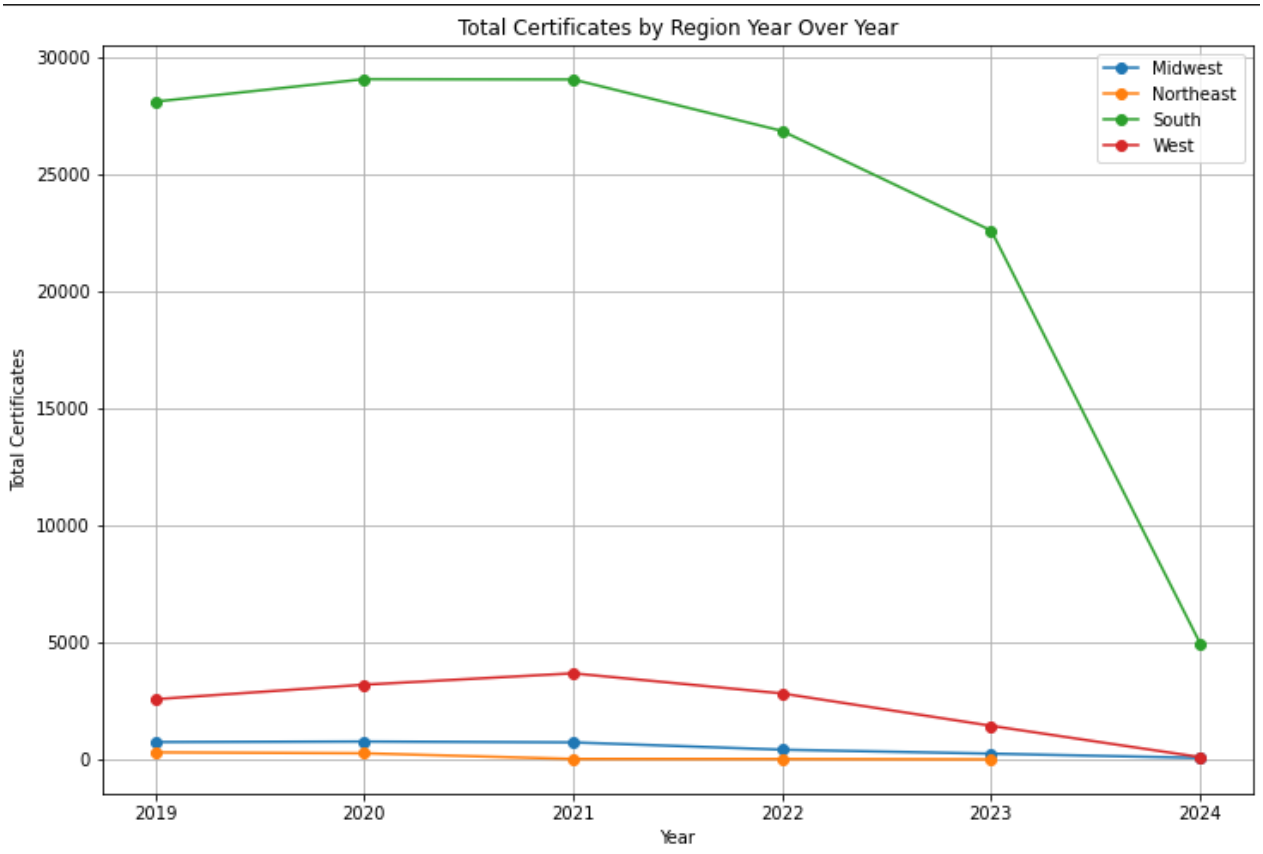
- The aggregated data from both datasets were merged on the state identifier to facilitate a combined analysis, allowing us to view the total premiums and contract costs side by side for each state.
- The regions were mapped into merged dataset for analysis.

This comprehensive approach to data preparation and analysis ensures that the datasets are clean, standardized, and ready for detailed exploratory data analysis and subsequent modeling. The following sections will delve deeper into exploratory data analysis, clustering, and predictive modeling to uncover actionable insights.

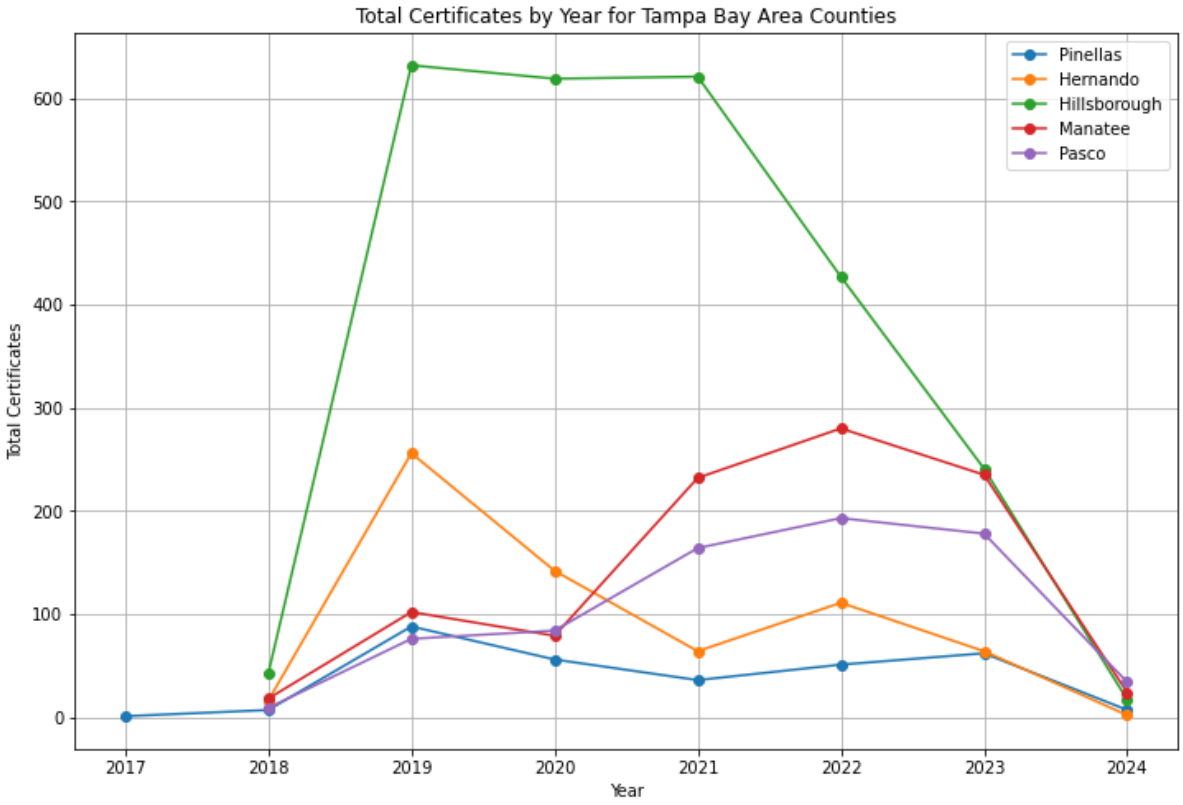
Data Visualization

Visualizing key variables and their relationships

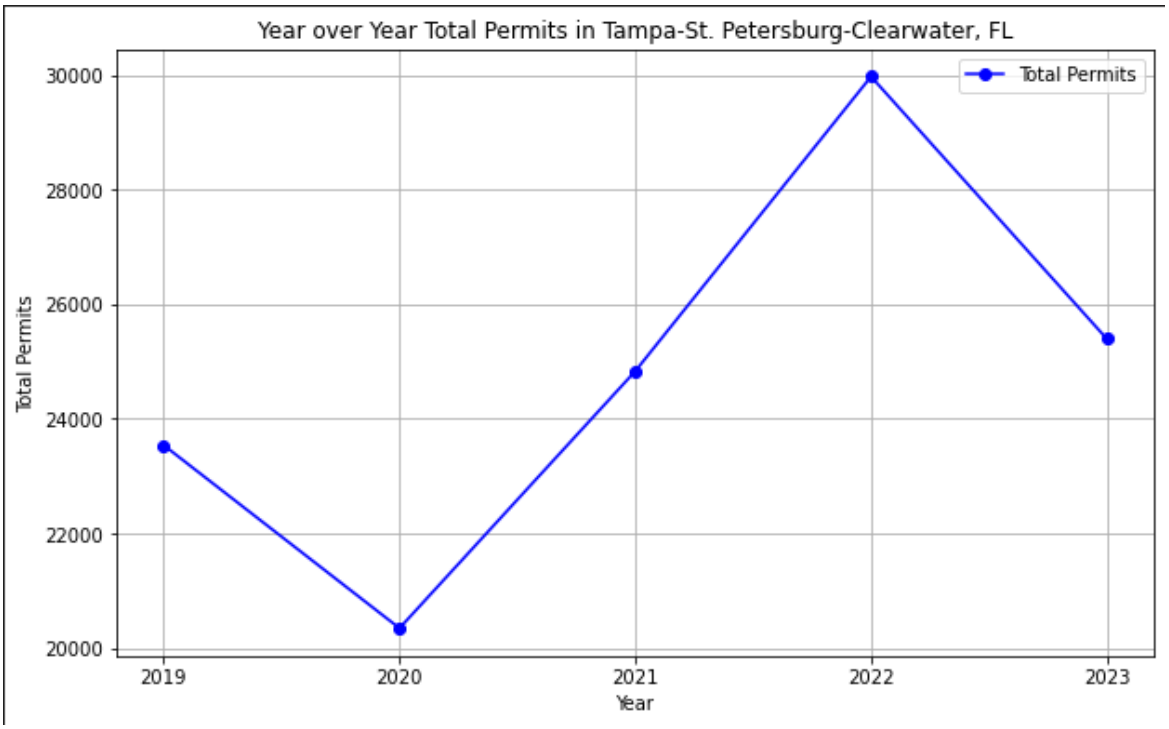
After combining the datasets to include regions by counts of certificates sold, the visualization below depicts that the South has traditionally been the region of focus. This graph shows a decline in policies sold.



Pivoting to our area of focus, Tampa Bay – Florida, within the initial stages of the analysis:

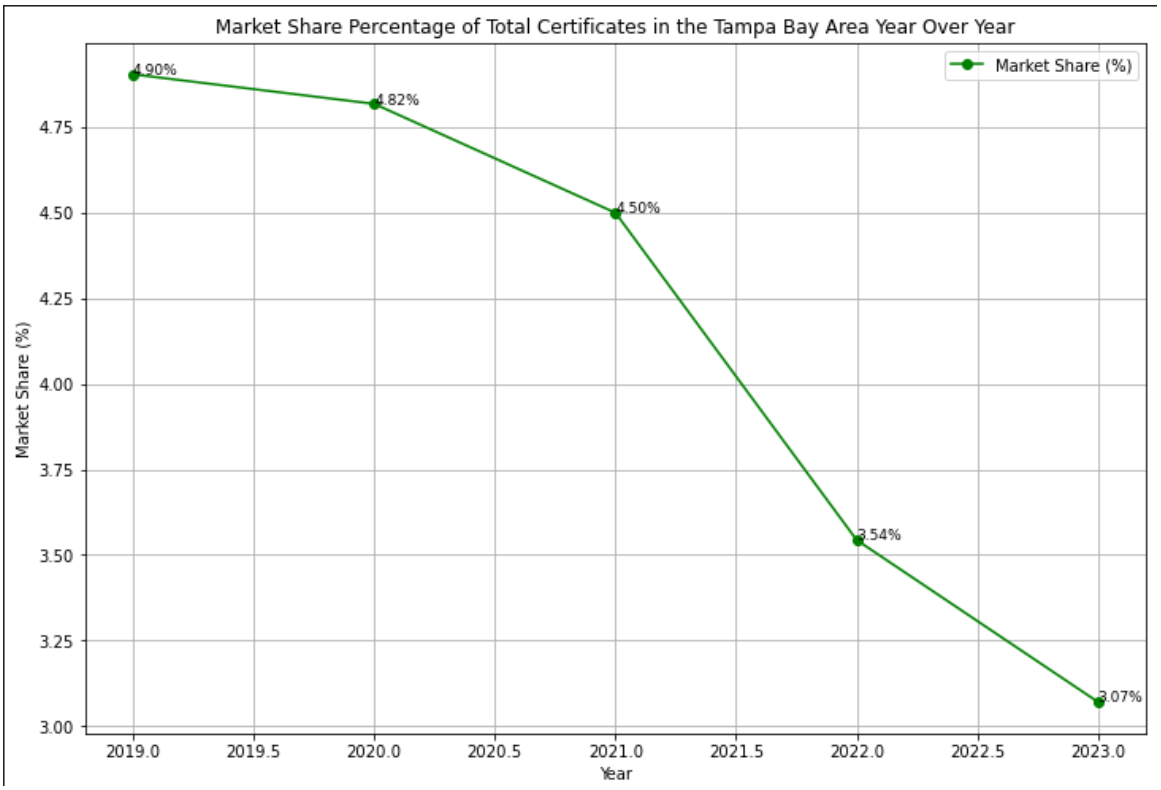


Showing Total permits for Tampa MSA:



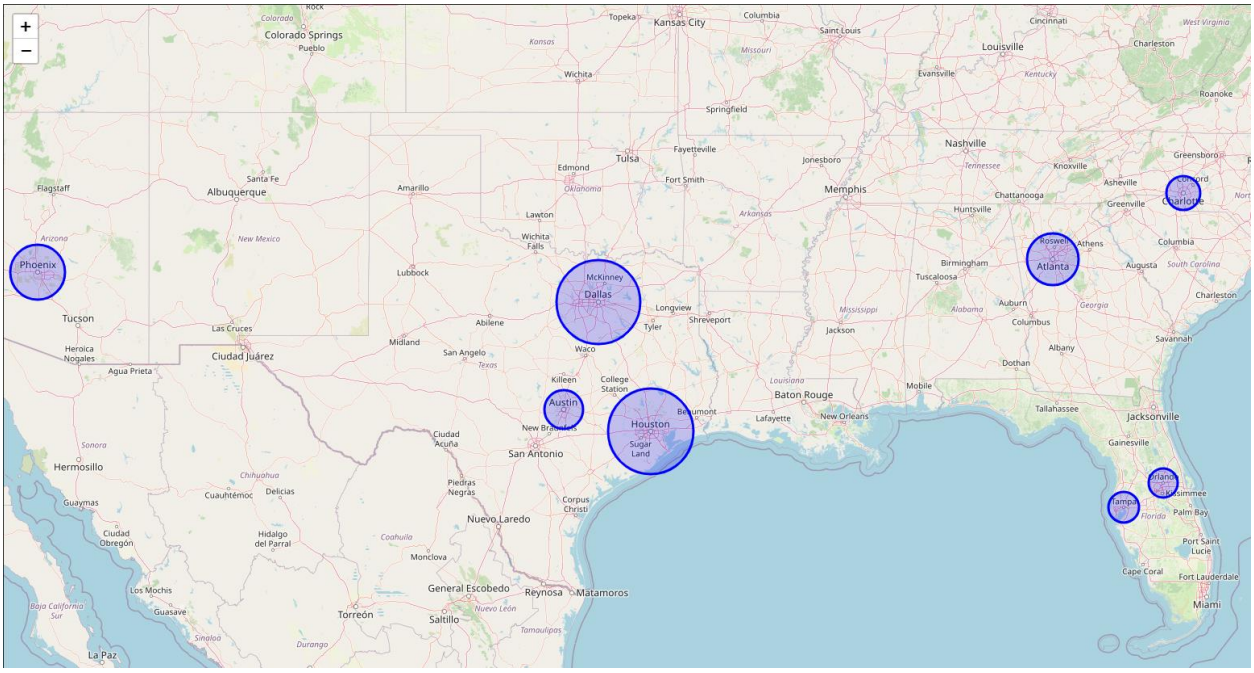
Methods and Analysis:

The analysis was generated to show MAV market share in the Tampa CBSA, From the graph below, you can see a steady decline, indicating a loss of market share.



Top List of Permits Pulled for New Housing:
In anticipation of generating a predictive tool to identify the next flywheel location best suited to MAV operations, these locations were identified using the beautiful soup python package to parse the census website and create a new data frame.

	CBSA	Location	Permit	Latitude	Longitude
0	26420	Houston, TX	52719.0	29.758938	-95.367697
1	19100	Dallas, TX	51996.0	32.776272	-96.796856
2	38060	Phoenix, AZ	34347.0	33.448437	-112.074141
3	12060	Atlanta, GA	31560.0	33.748992	-84.390264
4	12420	Austin, TX	24486.0	30.271129	-97.743700
5	16740	Charlotte, NC	20830.0	35.227209	-80.843083
6	45300	Tampa, FL	19305.0	27.947760	-82.458444
7	36740	Orlando, FL	17795.0	28.542111	-81.379030



Predictive Modeling: Using statistical and machine learning algorithms, I constructed models to forecast market potentials and identify key drivers of the sales flywheel.

- Logistic Regression: Developed a logistic regression model to predict the potential for successful market penetration, classifying regions as high or low potential based on their characteristics.
- Random Forest: A secondary model to possibly assess and compare results, providing a deeper insight into feature importance and model robustness.

Predictive Model Outcomes:

Logistic Regression Evaluation:
Accuracy: 1.0
Precision: 1.0
Recall: 1.0
F1-Score: 1.0
Confusion Matrix:
[[1 0]
 [0 1]]
Cross-Validation Scores: [1. 0. 0.5]
Average Cross-Validation Score: 0.5

Random Forest Evaluation:
Accuracy: 1.0
Precision: 1.0
Recall: 1.0
F1-Score: 1.0
Confusion Matrix:
[[1 0]
 [0 1]]

Cross-Validation Scores: [1. 0. 0.5]
Average Cross-Validation Score: 0.5

The predicted next area of flywheel generation is: Low Potential

Both models (Logistic Regression and Random Forest) show perfect accuracy, precision, recall, and F1-scores on the test set. However, this is misleading and typically indicative of overfitting, especially with small datasets.

The cross-validation scores reveal that the models do not generalize well, with scores varying significantly (1.0, 0.0, 0.5), leading to an average score of 0.5. This indicates that the model's performance is not stable across different splits of the data.

The high-test set performance and low cross-validation scores suggest overfitting. The models perform perfectly on the test set due to the small size and potentially non-representative nature of the test set.

The discrepancy between test set performance and cross-validation scores highlights the need for more data or better data sampling methods to ensure the models can generalize well to unseen data.

Next Steps for Improvement:

1. Increase Data Size: Collect more data to ensure the models have sufficient information to learn and generalize.
2. Handle Missing Data: Impute or address missing values in df_warranty_sales_cleaned to improve data quality.
3. Use Balanced Dataset: Ensure that the dataset used for training and evaluation is balanced in terms of the target variable to avoid biases.
4. Improve Cross-Validation: Use techniques like stratified cross-validation to ensure that each fold is representative of the overall dataset distribution.

Conclusion:

The analysis concluded that MAV's market share in the Tampa Bay area MSA has been in decline since 2019, necessitating strategic adjustments. By leveraging comprehensive data analysis, high-potential regions for the next generation of flywheel discovery were identified, including Houston, Dallas, and Austin, TX, among others. The findings suggest that the sales team in Texas is particularly well-positioned to implement the flywheel approach due to their established presence and understanding of the local market dynamics. Focusing on these high-potential regions and empowering the Texas sales team will be crucial for MAV to regain market share and drive sustained growth.

Assumptions:

- **Data Completeness and Accuracy**: It is assumed that all datasets used in this analysis are complete and accurate as extracted from their respective systems of record. This assumption underpins the reliability of the insights derived from the data. If the datasets contain significant gaps or inaccuracies, it could potentially skew the analysis results and lead to incorrect conclusions.
- **Consistent Data Collection Methods**: The analysis assumes that the data collection methods across different datasets are consistent. This includes the manner in which data points are

recorded, the frequency of data collection, and the definitions of various data fields. Inconsistencies in data collection methods could result in misalignment of data and affect the validity of the analysis.

- **Stable Market Conditions**: The predictive models developed assume that the market conditions will remain relatively stable in the short to medium term. Any sudden changes in economic factors, regulatory environments, or competitive landscapes could impact the accuracy of the models' predictions.
- **Representative Sample**: It is assumed that the datasets used are representative of the entire market. This includes if the sales data from MAV Corp and the permitting data from the U.S. Census Bureau adequately reflect broader market trends. Any biases in the sample could lead to misleading results.
- **Uniform Regional Dynamics**: The analysis assumes that the economic and demographic dynamics within each region are uniform enough to allow for meaningful aggregation and comparison. Variations within regions that are not accounted for could affect the accuracy of regional comparisons and predictions.
- **Static Customer Behavior**: The predictive models assume that customer behavior and preferences remain static over time. Significant shifts in customer preferences or buying behavior, influenced by factors such as technological advancements or cultural changes, could affect the models' accuracy.

By clearly stating these assumptions, I acknowledge the limitations of the analysis and set the context for interpreting the results. It also highlights areas where further data collection and validation might be necessary to strengthen the analysis.

Challenges and Limitations:

- **Data Scope and Availability**: The availability and scope of the datasets used might not capture all relevant variables or may be limited to a specific geographic region or time.
- **Historical Data Application**: Using historical data to predict future trends assumes that past patterns will continue, which can be problematic in volatile markets.
- **Generalizability**: Insights and patterns may not always hold true across different market segments or external conditions.
- **Data Quality and Integration**: Inconsistencies in how data is collected, processed, and stored can affect integration and overall analysis.
- **Data Integrity**: Variabilities in data quality could impact the accuracy of insights. Regular audits and validation checks will be implemented to mitigate this risk.
- **Adaptation to Regional Dynamics**: The diverse economic landscapes across different regions may influence the generalizability of the models. Tailored approaches will be developed to accommodate local market conditions.

Future Uses/Additional Applications:

The insights and models developed in this project have the potential to be applied beyond the Tampa Bay area, extending to other regions and industries. By adapting the models to account for local market dynamics, demographic trends, and economic conditions, MAV can identify new opportunities for market expansion and optimize resource allocation across various geographies. Additionally, the analytical framework used here can be tailored to different sectors, enabling businesses to enhance their sales strategies, improve customer retention, and predict high-growth areas. Continuous refinement and testing of the models in diverse settings will help in validating their robustness and scalability.

Recommendations:

Based on the analysis, the following actionable recommendations are proposed to enhance market penetration and streamline sales processes:

1. **Focus on High-Potential Regions**: Prioritize efforts in identified high-potential regions such as Houston, Dallas, and Austin. Allocate more resources and tailor marketing strategies to capitalize on these opportunities.
2. **Strengthen the Texas Sales Team**: Empower the Texas sales team with additional training, tools, and support to leverage their local market expertise. Encourage data-driven decision-making to maximize their effectiveness.
3. **Investigate Declining Market Share**: Conduct a thorough analysis to identify the factors contributing to the declining market share in Tampa Bay. Develop targeted interventions to address these issues.
4. **Enhance Data Collection and Quality**: Invest in improving data collection methods to ensure accuracy and completeness. Regularly update and validate datasets to support reliable analysis.
5. **Expand Predictive Modeling**: Continue developing and refining predictive models to identify emerging trends and opportunities. Use these insights to proactively adjust strategies and stay ahead of market shifts.
6. **Collaborate Across Teams**: Foster collaboration between sales, marketing, and data analytics teams to ensure a unified approach towards achieving business goals.

Implementation Plan:

To effectively implement the recommendations, the following plan outlines the necessary steps, resources, and timelines:

1. **Phase 1: Preparation (0-3 months)**
    - **Resource Allocation**: Identify and allocate necessary resources, including budget, personnel, and technology.
    - **Training and Development**: Conduct training sessions for the Texas sales team on the new flywheel strategy and data-driven decision-making.
    - **Data Enhancement**: Improve data collection methods and establish regular validation processes.
2. **Phase 2: Execution (4-12 months)**
    - **Market Focus**: Implement targeted marketing and sales strategies in high-potential regions such as Houston, Dallas, and Austin.
    - **Pilot Programs**: Launch pilot programs to test the refined sales flywheel model in the identified regions and adjust based on feedback and initial results.
    - **Ongoing Analysis**: Continuously monitor market trends and sales performance. Use predictive models to adjust strategies proactively.
3. **Phase 3: Evaluation and Scaling (13-24 months)**
    - **Performance Review**: Conduct a comprehensive review of the implementation outcomes. Assess the effectiveness of the strategies and the impact on market share.
    - **Scaling Success**: Expand successful strategies to other regions and industries. Tailor approaches based on specific market dynamics and opportunities.
    - **Feedback Loop**: Establish a continuous feedback loop to refine models, strategies, and processes based on real-time data and insights.

By following this implementation plan, MAV can strategically navigate the challenges and opportunities identified in the analysis, driving sustained growth and market penetration.

Ethical Assessment:

- **Data Privacy Compliance**: All data handling will strictly adhere to the General Data Protection Regulation (GDPR) and other relevant privacy regulations to protect customer information. This includes implementing robust data security measures, ensuring data anonymization where necessary, and regularly auditing data practices to prevent unauthorized access and data breaches. Transparency with customers about how their data is collected, stored, and used will be maintained, ensuring informed consent, and building trust.
- **Mitigation of Bias**: To ensure equitable results, measures will be implemented to identify and mitigate potential biases in data collection and analysis methodologies. This involves using diverse and representative datasets, applying fair and unbiased algorithms, and continuously monitoring for any discriminatory patterns. Regular reviews and updates to the models will be conducted to address any emerging biases. Additionally, promoting inclusivity in the data science team can provide diverse perspectives, further reducing the risk of bias in the analysis.

These ethical considerations are essential for maintaining the integrity of the analysis, protecting customer rights, and ensuring that the insights and recommendations are fair and beneficial to all stakeholders.

References:

- Kotler, P., & Keller, K. L. (2016). Marketing management (15th ed.). Pearson Education, Inc.
- Brown, B., & Bughin, J. (2019). Big data, analytics, and the future of marketing & sales. McKinsey & Company.
- HubSpot. (n.d.). The Flywheel. Retrieved May 2, 2024, from HubSpot