Customer Performance Analysis via Regression

CSCA 5622 Supervised Learning Final Project

You can find the full code, data, and documentation for this project on GitHub:
[CSCA-5622-Supervised-Learning-Final-Project (github.com/treinart)]
https://github.com/treinart/CSCA-5622-Supervised-Learning-Final-Project.git

---

Section 1: Project Overview

1.1: Purpose

This project uses supervised regression modeling to evaluate dealership customer performance at the customer level. The goal is to predict each customer's average total sales, combining labor and parts. I want to understand which customers tend to bring the most value and why. That includes how often they visit, what type of work they do, and how profitable those jobs are.

The insights from this model support real business decisions. Dealers can use the results to identify high-value customers and offer better support or incentives. They can also flag lower-performing accounts and re-evaluate pricing, promotions, or strategy.

This dataset reflects real dealership behavior. It was designed from scratch using operational business logic. That makes it suitable not just for academic use but also for deployment in real environments.

1.2: Project Dependencies and Environment

This project requires the following Python packages:

- Pandas
- Numpy
- Matplotlib
- Seaborn
- Scikit-learn
- Scipy
- xgboost

If running this notebook or analysis for the first time, please ensure all packages are installed in your Jupyter or Python environment. To install everything at once, use:

```
!pip install pandas numpy matplotlib seaborn scikit-learn scipy xgboost
```

(Run this command in a Jupyter code cell, or drop the ! and run it in your terminal.)

If you see any import errors, check that your notebook kernel matches your intended Python environment. To verify which Python is being used, run:

```
import sys

print(sys.executable)
```

If all else fails (e.g., with xgboost), run:

<full_path_from_above> -m pip install xgboost

This was the only way I was successful installing XGBoost.

This will guarantee that XGBoost is installed in the environment Jupyter is using.

Restart the Jupyter kernel after installing.

In a code cell, try:

from xgboost import XGBRegressor

If you do not get an error, installation was successful.

Section 2: Data Description and Preparation

2.1: Data Source and Origin

The dataset was generated using a custom Python script called generate_invoice_data.py. This script simulates dealership invoice activity using realistic rules and business assumptions. It does not include any actual customer names, dollar amounts, or internal records. The output file is anonymized_invoice_data.csv.

This approach protects confidentiality but keeps the structure and behavior of real dealership invoices. The data generation process is explained in full in the attached document, “Generate Anonymized Data Story.”

2.2: Data Description

The file contains 45,514 rows and 15 columns. Each row represents a single invoice across 387 unique customers. The data is stored in a flat table.

Columns include a mix of numeric and categorical fields. Examples:

- LaborBilled$ and PartsSales$ represent billed dollar amounts
- HoursWorked and HoursBilled track technician time
- ROtype, Dept, and Location categorize the job and where it occurred
- CustNo and CustName track the customer

This structure supports downstream aggregation to the customer level, which is the focus of this model.

2.3: Data Loading

The analysis begins by loading the generated CSV file into a pandas DataFrame with:

df = pd.read_csv("anonymized_invoice_data.csv")

This step makes the full invoice-level dataset available for cleaning, exploration, and modeling.

Note: The file anonymized_invoice_data.csv must be located in the same directory as the Jupyter notebook (.ipynb file) to ensure it loads correctly.

If storing the data elsewhere, update the pd.read_csv() path accordingly.

2.4: Data Cleaning

The data was designed to be clean. The generator script does not allow missing values or illogical values like negative dollar amounts. However, I still verified the dataset programming to confirm that it is ready for modeling.

Cleaning Steps Performed

1. **Preview the first few rows**
   Purpose: Visually check column names, order, and a small sample of data for sanity.

   Code/Output:
   df.head()

```
# Load the invoice-level dataset from CSV using pandas
df = pd.read_csv("anonymized_invoice_data.csv")   # This loads the entire file into a DataFrame called df

# Show the first five rows to confirm successful loading and review column names and sample data
df.head()
```

	Location	Whse	InvoiceNo	InvDate	CustNo	CustName	ROtype	Dept	HoursWorked	HoursBilled	LaborBilled$	LaborWorked$	PartsSales$	PartsCost$	InvC
0	Dallas	DL2	100000	05/12/2025	JJQ9K	Goldstar Associates	COUNTER	30	0.00	0.00	0.00	0.00	458.56	375.65	
1	Green Bay	GB1	100001	08/25/2022	JJQ9K	Goldstar Associates	TRAILER	40	2.44	1.94	604.28	292.80	457.04	252.25	
2	Dallas	DL1	100002	09/21/2022	JJQ9K	Goldstar Associates	RESALE	40	6.27	7.01	1870.92	1097.25	966.72	757.59	
3	Chicago	CH1	100003	04/29/2025	JJQ9K	Goldstar Associates	RESALE	20	5.40	7.40	1940.86	945.00	272.45	240.13	
4	Green Bay	GB4	100004	10/26/2023	JJQ9K	Goldstar Associates	TRAILER	50	4.68	5.23	1270.51	561.60	762.96	606.50	

2. **Check number of rows and columns**
   Purpose: Confirm the data size matches expectations and contains all fields.

   Code/Output:
   df.shape

```
# Display the number of rows and columns in the dataset
df.shape   # Output: (number of rows, number of columns)
print("Rows, columns:", df.shape)
```

```
Rows, columns: (45514, 15)
```

3. **Check number of unique customers**
   Purpose: Determine how many distinct customers are represented in the data, as opposed to total invoice count. This helps clarify the analysis scope and ensures that aggregation or segmentation steps are based on accurate customer-level counts.

Code/Output:
print("Number of unique customers:", df["CustNo"].nunique())

```
# Print the number of unique customers in the dataset.
# This counts how many distinct customer IDs (CustNo) appear in the invoice-level data.
print("Number of unique customers:", df["CustNo"].nunique())
```

```
Number of unique customers: 387
```

4. **Check for missing values**
   Purpose: Ensure there are no nulls in any column.
   All columns showed 0 missing values.

   Code/Output:
   df.isnull().sum()

```
# Check for missing values in each column to confirm data integrity
df.isnull().sum()    # Expect all columns to show 0 missing values
print("\nMissing values:\n", df.isnull().sum())
```

```
Missing values:
 Location        0
Whse             0
InvoiceNo        0
InvDate          0
CustNo           0
CustName         0
ROtype           0
Dept             0
HoursWorked      0
HoursBilled      0
LaborBilled$     0
LaborWorked$     0
PartsSales$      0
PartsCost$       0
InvCycleDays     0
dtype: int64
```

5. **Review column data types**
   Purpose: Make sure every field is correctly typed (e.g., dollars as float, dates as string or datetime, categories as object).

   Code/Output:
   df.dtypes

```
# Review the data types (e.g., float, int, object) for each column
df.dtypes    # This helps identify if any columns are mis-typed
print("\nColumn types:\n", df.dtypes)
```

```
Column types:
 Location         object
Whse              object
InvoiceNo          int64
InvDate           object
CustNo            object
CustName          object
ROtype            object
Dept               int64
HoursWorked      float64
HoursBilled      float64
LaborBilled$     float64
LaborWorked$     float64
PartsSales$      float64
PartsCost$       float64
InvCycleDays     float64
dtype: object
```

6. **Review summary statistics**
   Purpose: Check for any out-of-bounds or illogical values (e.g., negative sales, negative hours).

   Code/Output:
   df.describe()

```
# Show summary statistics (mean, std, min, max, quartiles) for all numeric columns
df.describe()
```

	InvoiceNo	Dept	HoursWorked	HoursBilled	LaborBilled$	LaborWorked$	PartsSales$	PartsCost$	InvCycleDays
count	45514.000000	45514.000000	45514.000000	45514.000000	45514.000000	45514.000000	45514.000000	45514.000000	45514.000000
mean	122756.500000	30.054489	1.943942	2.473220	902.099548	308.744873	828.727662	534.102356	7.449488
std	13138.904413	14.095500	2.400496	3.195262	1342.479969	388.871755	520.255687	298.860757	4.105931
min	100000.000000	10.000000	0.000000	0.000000	0.000000	0.000000	-299.360000	-127.570000	1.000000
25%	111378.250000	20.000000	0.000000	0.000000	0.000000	0.000000	415.047500	281.980000	4.100000
50%	122756.500000	30.000000	0.000000	0.000000	0.000000	0.000000	776.615000	527.845000	7.300000
75%	134134.750000	40.000000	4.110000	4.940000	1614.620000	631.750000	1158.635000	776.700000	10.600000
max	145513.000000	50.000000	6.900000	14.490000	13222.220000	1207.500000	2975.650000	1251.410000	15.000000

Results

All data integrity checks passed:

- No missing values were found in any column.
- All columns have the correct data types.
- There are no negative or obviously illogical values in numeric fields.
- The number of rows and columns matches the design of the generator script.

Because the data was programmatically generated to be clean, no manual cleaning was necessary. These checks confirm the dataset is ready for aggregation, EDA, and modeling.

In a real-world project, we would expect to see and handle missing values, mis-typed columns, or invalid entries at this stage. Here, the only anomalies observed were in rare outlier values, which are expected for a business of this size and scope.

2.5: Cleaning Summary

There were no missing values in the file. All data types were correct. Nothing needed to be dropped or imputed. The distributions for labor and parts looked normal for a service business.

No changes were made to the original file. However, later steps like feature engineering may include capping or transforming some fields. For example, may limit extreme values in efficiency or gross margin to avoid distorting model results.

The script that generated this dataset, along with the documentation, gives us confidence in its structure and reliability.

2.6: Visual Checks

In this stage, I use visualizations to quickly assess the distribution and integrity of key numeric fields.
Our goals are to:

- Understand where most jobs fall in terms of size and value
- Identify any outliers or unusual values that may require attention
- Check for skew or irregularities that could affect modeling or require feature transformation
- Confirm that the data is clean and representative of realistic dealership operations

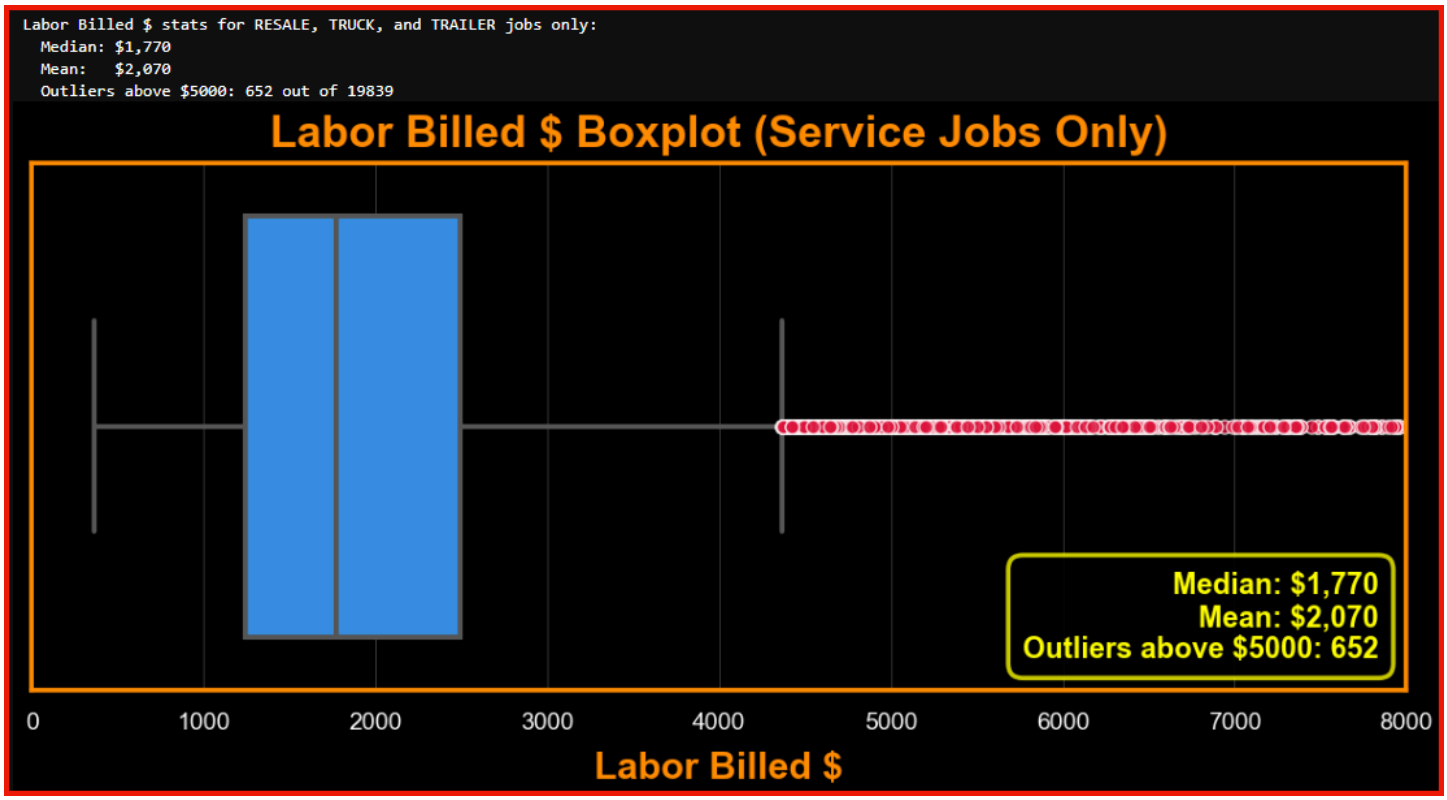
Visual inspection is a critical step in any applied data science project. It ensures the data "makes sense" before further analysis and provides an intuitive feel for business patterns that may not be obvious in summary statistics alone.

2.7: Boxplot: Labor Billed $ (All Service Jobs)

This boxplot summarizes the distribution of Labor Billed $ across all service jobs, including RESALE, TRUCK, and TRAILER work. The median labor billed for these jobs is $1,770 and the mean is $2,070. There are 652 invoices above $5,000, which indicates a substantial number of high-value outliers.

However, the plot itself provides limited business insight. The majority of invoices cluster tightly below $4,000, while the outliers stretch far beyond that range, making it difficult to see meaningful differences or trends in the typical job. The high concentration of routine jobs combined with a long right tail of extreme values flattens the visual and obscures the operational patterns management needs to see.

This experience shows that standard boxplots are not always optimal for every dataset. For these reasons, I broke the analysis down further, segmenting by ROtype and by location. This approach uncovers where the highest labor sales occur and how job type or branch impacts billed amounts. The segmented plots provide more actionable insights for management and help guide decisions on pricing, staffing, and sales strategy.



2.8: Boxplot: Labor Billed $ by ROtype

The boxplot for Labor Billed $ by ROtype compares RESALE, TRAILER, and TRUCK jobs, providing a clear picture of how billed labor varies by repair type.

RESALE jobs account for the largest share of service work, with more than 13,000 jobs in the data. The median labor billed for RESALE is $1,974, and the mean is $2,272. These jobs also show the highest single-invoice value, with a maximum of $13,222. The wide spread and presence of high-value outliers suggest a mix of routine service jobs and more complex, high-dollar repairs.

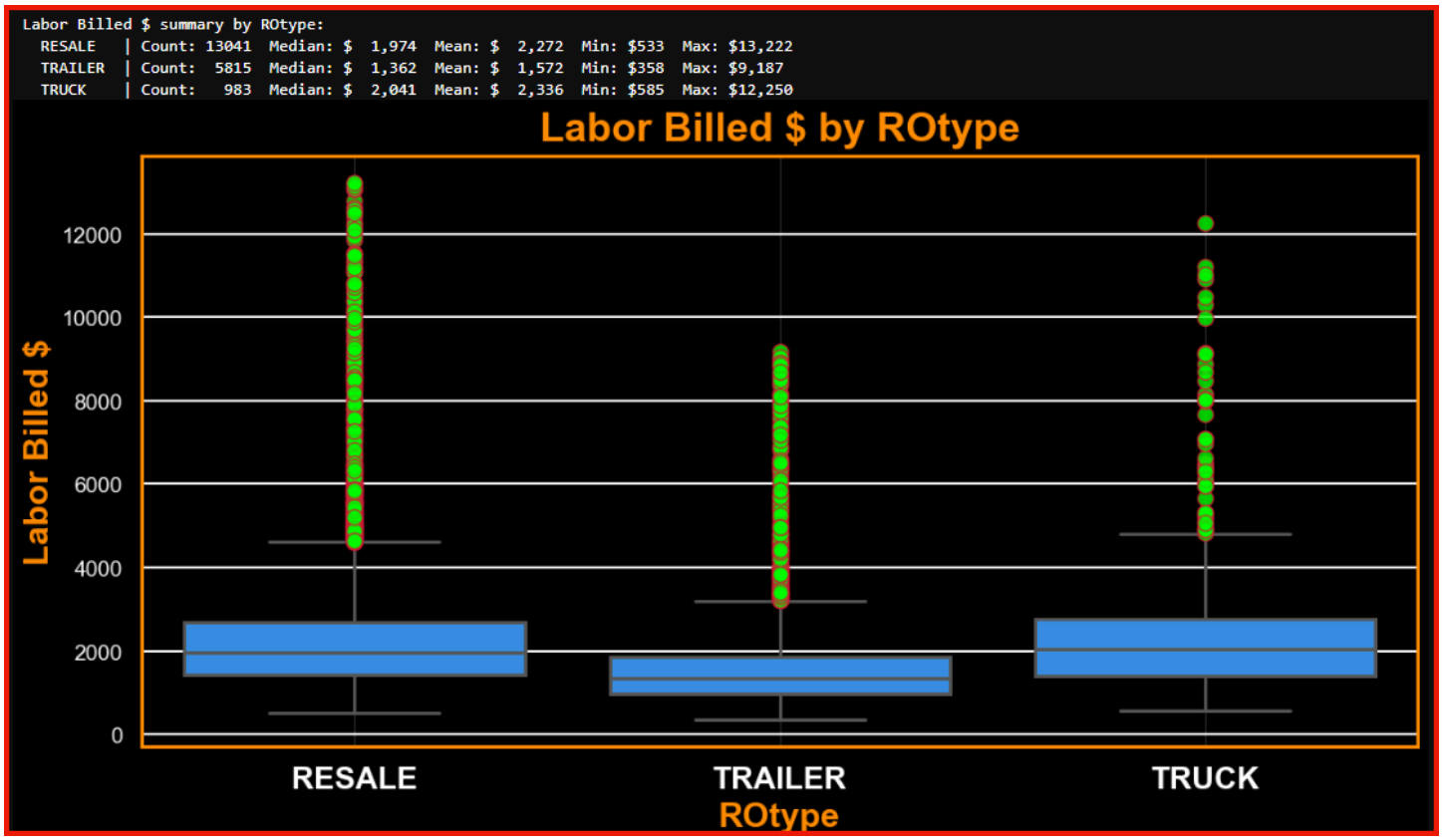
TRAILER jobs make up the second largest category, with nearly 6,000 jobs. The typical TRAILER job bills less labor, with a median of $1,362 and a mean of $1,572. Although some outlier jobs reach nearly $9,200, most TRAILER jobs are clustered in a narrower, lower range. This aligns with industry expectations, as trailer repairs are often less labor-intensive but may be completed more efficiently.

TRUCK jobs are the smallest group by count, at under 1,000 jobs, but they have the highest median labor billed ($2,041) and the highest mean ($2,336) of the three groups. The distribution is similar to RESALE, with a long tail of high-value outliers and a maximum invoice of $12,250. TRUCK jobs, while less frequent, tend to be larger and more complex, driving higher labor sales per invoice.

The plot and data highlight how labor billed varies not only with job volume, but also with job type and complexity.

Management can use this segmentation to understand where the highest value work is occurring and to set goals for improving the mix of high-value jobs or raising the average value of lower-billed categories. In particular, even small improvements in the average labor billed for TRAILER jobs or growth in TRUCK jobs could have a significant impact on overall revenue.

RESALE work brings the most volume, but TRUCK jobs bring the highest labor billed per job. Outliers and long tails are present in every category, showing potential for both risk and reward when managing the service mix.



2.9: Boxplot: Labor Billed $ by Location

The grouped boxplot of Labor Billed $ by location highlights significant differences in sales performance across branches.

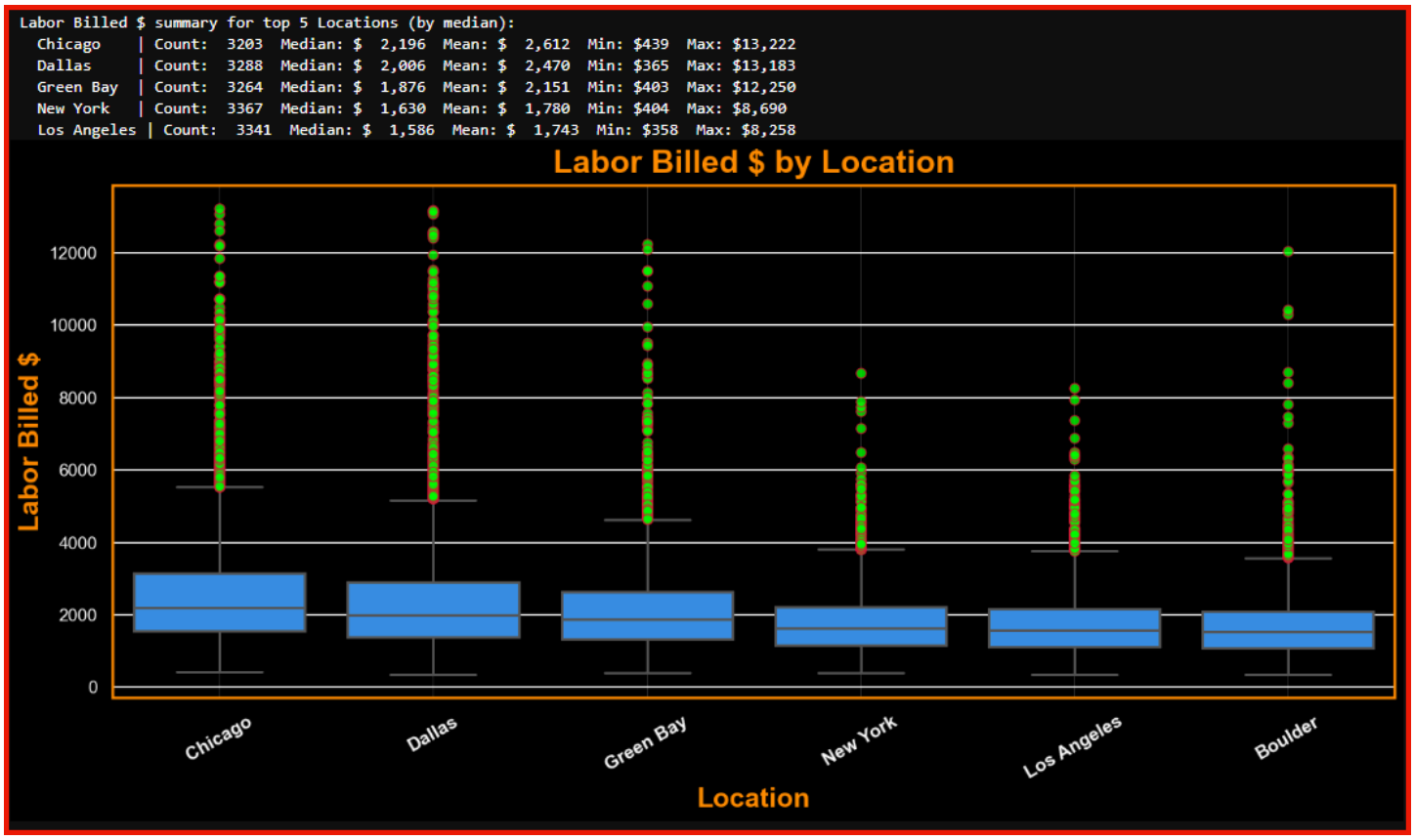
- **Chicago, Dallas, and Green Bay** show the highest median labor billed per invoice, all with medians above $1,800.
- **New York and Los Angeles** handle similar invoice volume, but have noticeably lower median and mean labor sales.

For example, **Los Angeles processed 3,341 invoices**, which is 4.3% more than Chicago's 3,203. However, the **average labor billed per invoice in LA is $1,743, about $610 lower than Chicago's $2,353 mean.** If LA could close even half that gap (a $305 increase per invoice), it would add over $1 million in annual labor sales, based on their current volume ($305 × 3,341 = $1,019,105).

This analysis shows that the highest revenue opportunities may not come from more invoices alone, but from raising the average value per job. The boxplot, combined with summary stats, makes these opportunities immediately visible.

Management can use this information to investigate why some locations outperform others whether due to pricing, job mix, efficiency, or local market factors and set targeted goals for closing the gap.

Even modest improvements in underperforming locations can drive major revenue gains. This type of customer and branch segmentation is central to a data-driven business strategy.



2.10: Boxplot: Total Sales per Invoice for Top 10 Customers

This boxplot compares the distribution of total sales per invoice for the company's ten highest-revenue customers. Each box summarizes the sales spread for a single customer, showing both typical invoice values and the presence of exceptionally large or small transactions.

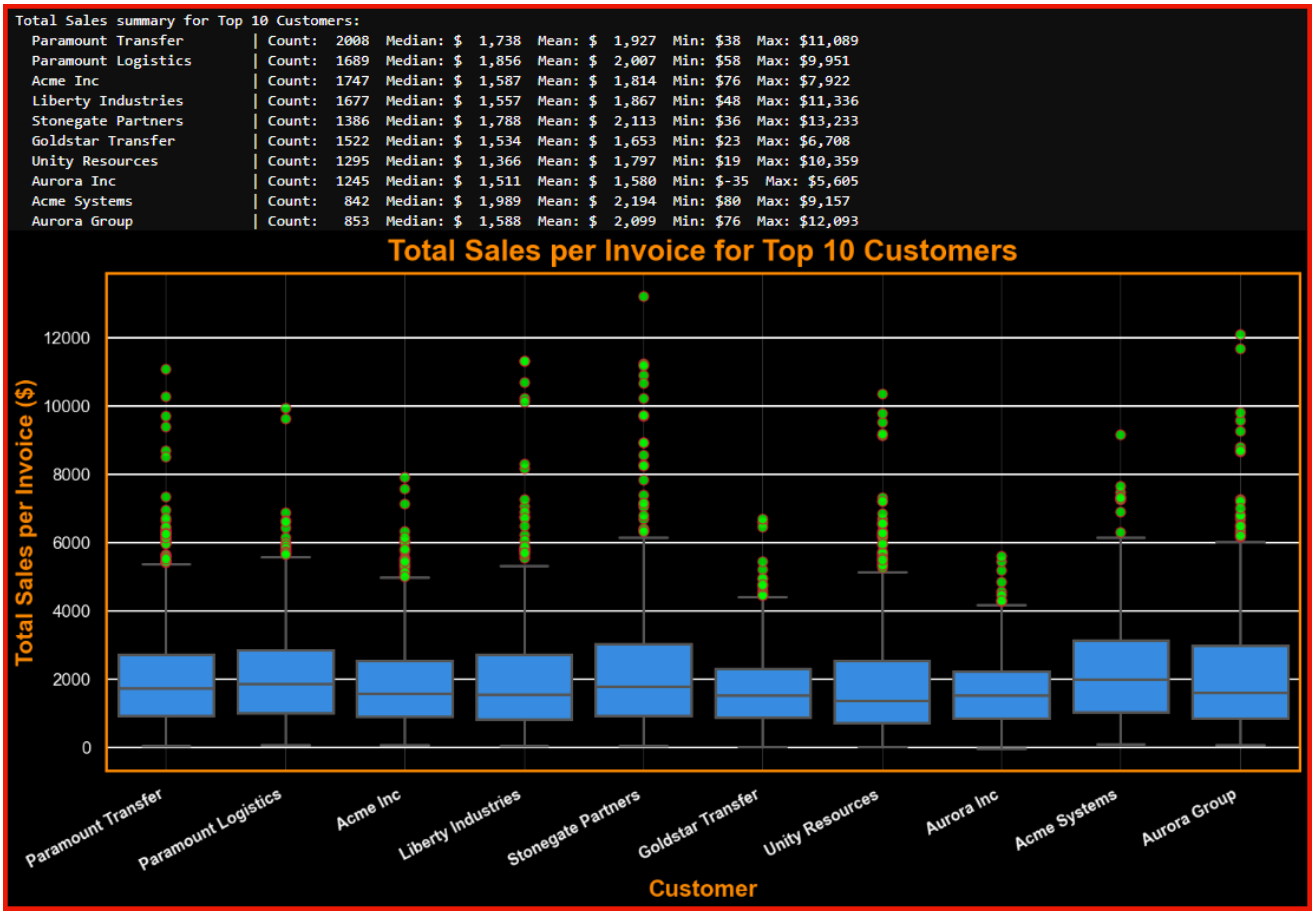
The analysis shows that most top customers have a relatively consistent invoice size, with medians typically ranging from about $1,200 to $2,000. However, there are important differences in both volume and invoice value. For example, Goldstar Associates not only processes the highest number of invoices but also generates some of the highest individual sales, with outliers well above $5,000. Other customers, such as Aurora Logistics and Pioneer Group, also show strong sales but have narrower spreads and fewer extreme outliers.

The summary statistics confirm these visual patterns. While all top customers drive substantial revenue, some rely on frequent, moderate-sized invoices, while others occasionally post very high-value sales. This diversity in invoice profiles highlights different business relationships and sales strategies at play.

From a management perspective, this analysis is valuable for both retention and growth planning. Large, consistent customers are essential for stable revenue. However, identifying what drives high-value outliers for specific accounts could help replicate that success across more customers. Additionally, if even a few moderate-value customers increase their average invoice size, the impact on total sales could be significant.

Understanding both the sales consistency and outlier behavior of top customers enables more targeted engagement, supports risk management, and creates actionable opportunities for revenue growth.

The top ten customers display a healthy diversity in both the number of jobs and the average value per job. Some focus on high volume, others on high-value transactions, and several accounts show significant outlier activity. These differences highlight the need for tailored account management and suggest opportunities for targeted growth by either increasing volume with high-average customers or raising invoice size for high-volume accounts. Negative invoice values should be investigated for potential corrections.



2.11: Distribution of Labor Billed $ (Service Jobs Only)

After examining labor sales with box plots, I turned to histograms to get a clearer view of how Labor Billed $ is distributed for all service jobs (RESALE, TRUCK, and TRAILER). The box plots allowed us to compare categories, but did not reveal much about the underlying frequency of invoice amounts.

The histogram below shows that the vast majority of service invoices cluster between $1,000 and $3,000, with a median value of $1,770 and a mean of $2,070. There are also 652 invoices above $5,000, producing a long right tail that pulls the mean above the median. This kind of right-skewed distribution is typical in service businesses, where a few complex jobs can be much larger than routine work.

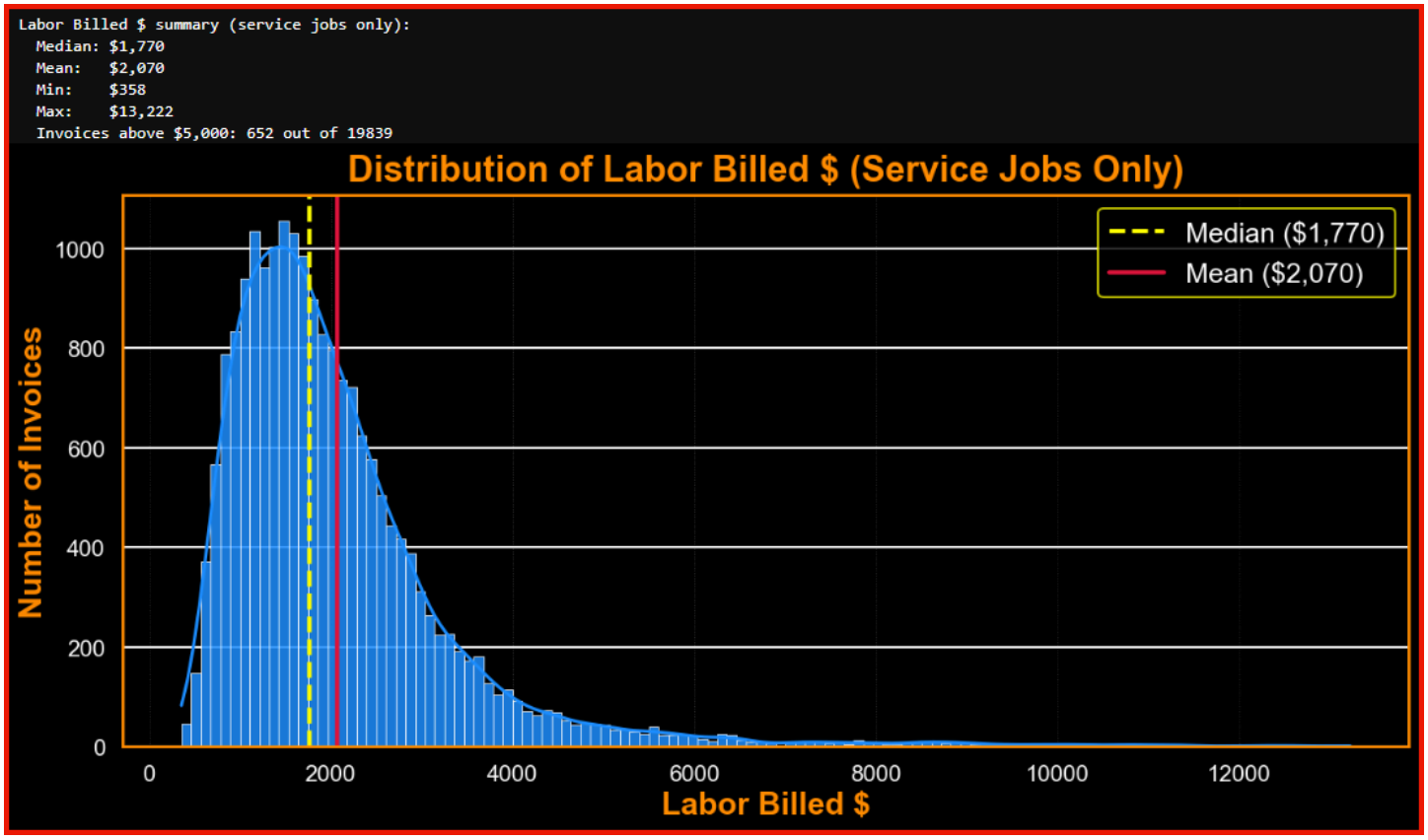
While this first histogram helps illustrate the spread of labor sales, it is still dominated by high-value outliers. Most invoices fall within a much tighter range, and the presence of a few large jobs compresses the rest of the plot, making it difficult to see the most common invoice values.

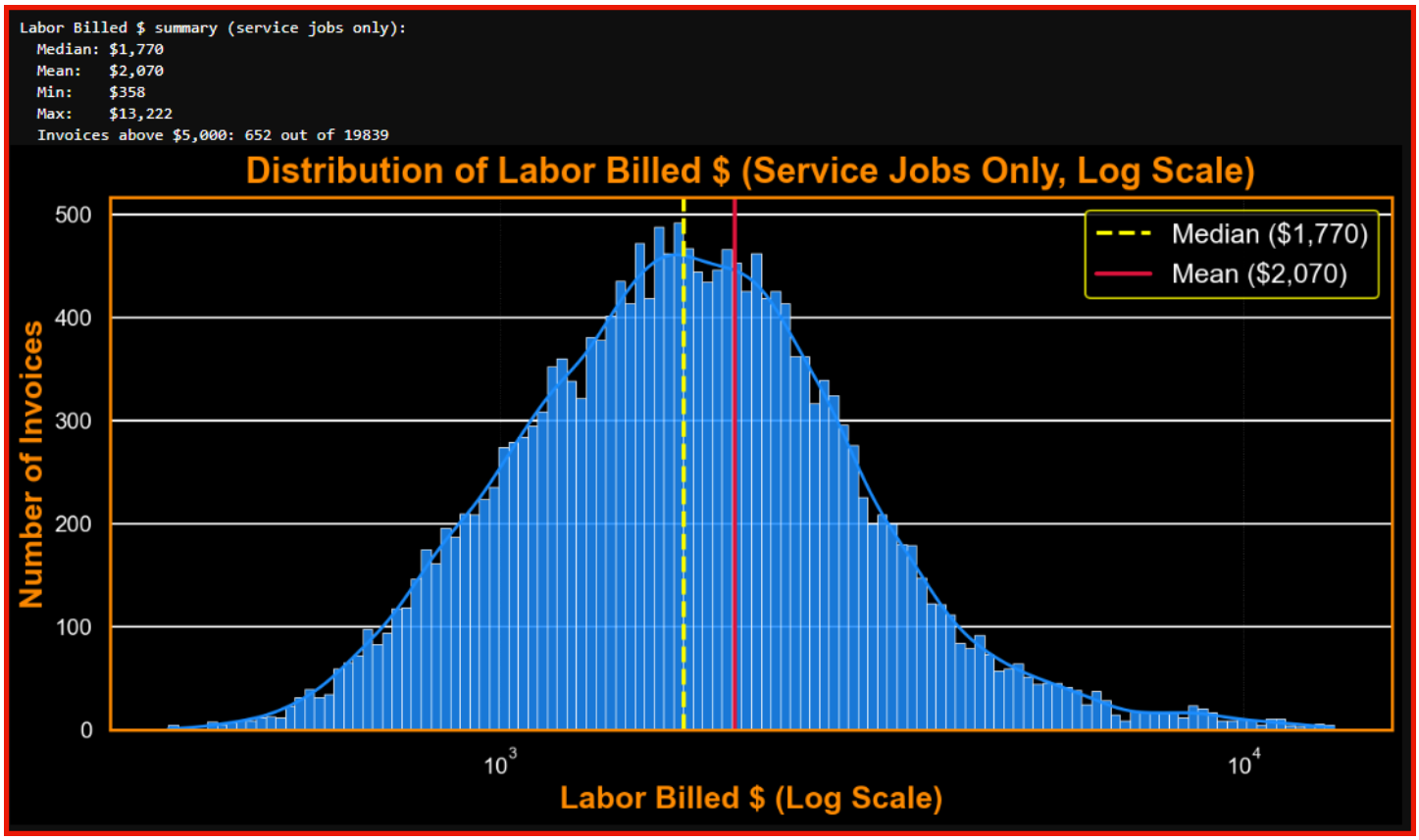
To solve this, I plotted the same data with a logarithmic x-axis. The log scale makes both typical and high-value jobs visible at the same time, revealing that the distribution is actually much more concentrated around the $1,500–$2,500 range than it appears in the standard plot. This allows management to better identify revenue patterns, see where most jobs fall, and spot outliers more easily.

Key insights:

- Most labor sales fall between $1,000 and $3,000, with few extremely large jobs driving the right tail.
- The distribution is right-skewed, which means a few big jobs have a disproportionate impact on the average.
- Using a log scale helps reveal the true density of routine invoices and makes the outliers visually clear.

Visualizing labor billed with both regular and log histograms gives managers a fuller understanding of typical sales versus exceptional jobs, allowing for smarter benchmarking, forecasting, and risk management.



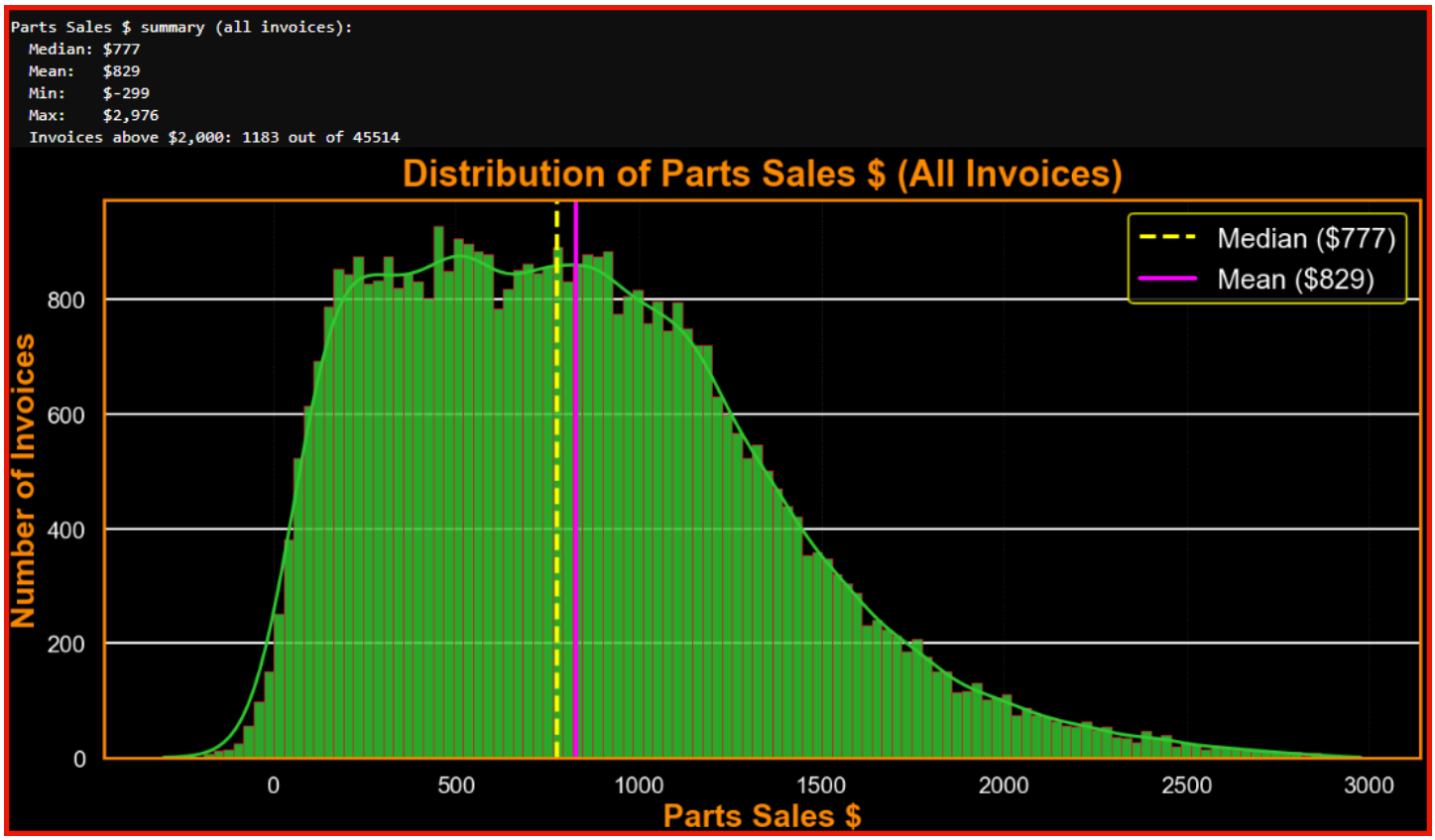


2.12: Distribution of Parts Sales $ (All Invoices)

This histogram shows the distribution of *Parts Sales $* for all invoices in the dataset, including both parts department (counter sales) and service department transactions. The majority of invoices fall between $0 and $1,500, with a median of $777 and a mean of $829. The right-skewed shape and long upper tail indicate that while most transactions are routine, there are a meaningful number of large parts invoices. 1,183 invoices are above $2,000, representing higher-value sales or major repair jobs.

The plot also reveals a small number of negative values, most likely representing customer returns or credits, which are typical in real dealership data. Overall, the distribution is less skewed than labor sales, but still demonstrates that a minority of high-value invoices contribute significantly to total parts revenue.

This distribution suggests that while most daily parts activity is at moderate invoice values, outliers (large jobs or bulk sales) play a meaningful role in the dealership's revenue. Management should be mindful of both everyday sales volume and the impact of retaining customers with high parts spend.



The initial EDA focused on invoice-level insights, visualizing the distribution and behavior of key variables such as LaborBilled$ and PartsSales$. While this was useful for spotting data quality issues and understanding invoice-level dynamics, our modeling goal is to analyze performance at the customer level.

To do this, I need to aggregate the invoice data to create a customer-level table. Each row in this dataset summarizes a unique customer's overall sales, number of visits, average efficiency, and other performance indicators. The following section explores this aggregated dataset to uncover the customer attributes most strongly linked to high average sales, and to identify trends that will inform the regression modeling.

Section 3: Exploratory Data Analysis (EDA)

3.1: EDA Overview

This section analyzes the customer-level dataset, which was created by aggregating the original invoice data from anonymized_invoice_data.csv. Each row in the customer-level table represents a unique customer and summarizes their historical performance across all invoices. The objective of this analysis is to identify trends, relationships, and potential predictors of average customer sales, which will inform our regression modeling.

3.2: Data Aggregation and Structure

Our business objective is to understand what drives high-value customers and predict their average total sales. To do this, I aggregated the invoice data to create a customer-level table. Each row in this table represents a unique customer and summarizes key performance metrics:

- **AvgTotalSalesPerInvoice**: The average of total sales per invoice (labor plus parts)

- **TotalInvoices**: The number of invoices linked to each customer
- **AvgLaborGM**: Average labor gross margin percentage
- **AvgPartsGM**: Average parts gross margin percentage
- **AvgEfficiency**: Average efficiency (hours billed divided by hours worked)
- **MostCommonROtype**: The most frequent job type (e.g., RESALE, COUNTER)
- **MostCommonDept**: The department most often used
- **MostCommonLocation**: The business location most frequently associated with the customer

There are 387 unique customers in the dataset. The aggregated table includes eight main features for each customer, in addition to their customer name and ID.

This step ensures our modeling and analysis reflect customer performance, not just individual transactions.

Aggregation Code:

```
# Calculate total sales for each invoice
df["TotalSales$"] = df["LaborBilled$"] + df["PartsSales$"]

# Calculate labor and parts gross margin percent (as a decimal)
df["LaborGM%"] = 1 - (df["LaborWorked$"] / df["LaborBilled$"])
df["PartsGM%"] = 1 - (df["PartsCost$"] / df["PartsSales$"])

# Calculate efficiency (hours billed divided by hours worked)
df["Efficiency"] = df["HoursBilled"] / df["HoursWorked"]

# Group by customer to create customer-level features
customer_df = df.groupby(["CustNo", "CustName"]).agg(
    AvgTotalSalesPerInvoice=("TotalSales$", "mean"),
    TotalInvoices=("InvoiceNo", "count"),
    AvgLaborGM=("LaborGM%", "mean"),
    AvgPartsGM=("PartsGM%", "mean"),
    AvgEfficiency=("Efficiency", "mean"),
    MostCommonROtype=("ROtype", lambda x: x.mode()[0] if not x.mode().empty else 'UNKNOWN'),
    MostCommonDept=("Dept", lambda x: x.mode()[0] if not x.mode().empty else -1),
    MostCommonLocation=("Location", lambda x: x.mode()[0] if not x.mode().empty else 'UNKNOWN')
).reset_index()

customer_df.head()
```

	CustNo	CustName	AvgTotalSalesPerInvoice	TotalInvoices	AvgLaborGM	AvgPartsGM	AvgEfficiency	MostCommonROtype	MostCommonDept	MostCommonLocatio
0	016ZZ	Goldstar Solutions	2005.883652	742	0.658874	0.345703	1.370037	COUNTER	20	Los Angele
1	04CDPQ	Summit Partners	1705.610000	23	0.574704	0.223554	1.024120	COUNTER	10	Green Ba
2	050E6	Evergreen Systems	2176.141463	82	0.611540	0.232640	1.076672	RESALE	40	Boulde
3	0AOBT3	Unity Logistics	1369.023846	13	0.592245	0.323652	1.501356	COUNTER	20	Dalla
4	0KB68	Liberty Solutions	1885.920909	11	0.756308	0.335894	1.298439	COUNTER	40	Green Ba

---

3.3: Customer-Level EDA

With the customer-level dataset created, I next examine the distribution and summary statistics for each feature. This gives a sense of how customer performance varies, highlights outliers, and identifies patterns that may be relevant for modeling.

The customer-level table includes 387 unique customers, each described by total sales per invoice, total invoices, average gross margins, efficiency, and categorical attributes.

- The average customer has about 118 invoices, but the range is wide from 1 up to over 2,000 invoices per customer.
- Average total sales per invoice span from $86 to $5,563, with a mean of $1,729 and a median of $1,641. The distribution is right-skewed, with a handful of high-volume customers.
- Labor and parts gross margins mostly fall between 0.55 and 0.66, indicating stable profitability for most customers, though some show higher or lower margins.
- Efficiency is centered around 1.2, with most customers clustered close to that value, but a few have much higher or lower efficiency rates.
- MostCommonDept ranges from 10 to 50, matching business department codes.

These summary statistics confirm that our aggregated dataset covers a wide range of customer activity and profitability. The spread and skew in key metrics will be important to keep in mind as we move to feature visualization and correlation analysis.

3.4: Distribution of Average Total Sales per Customer

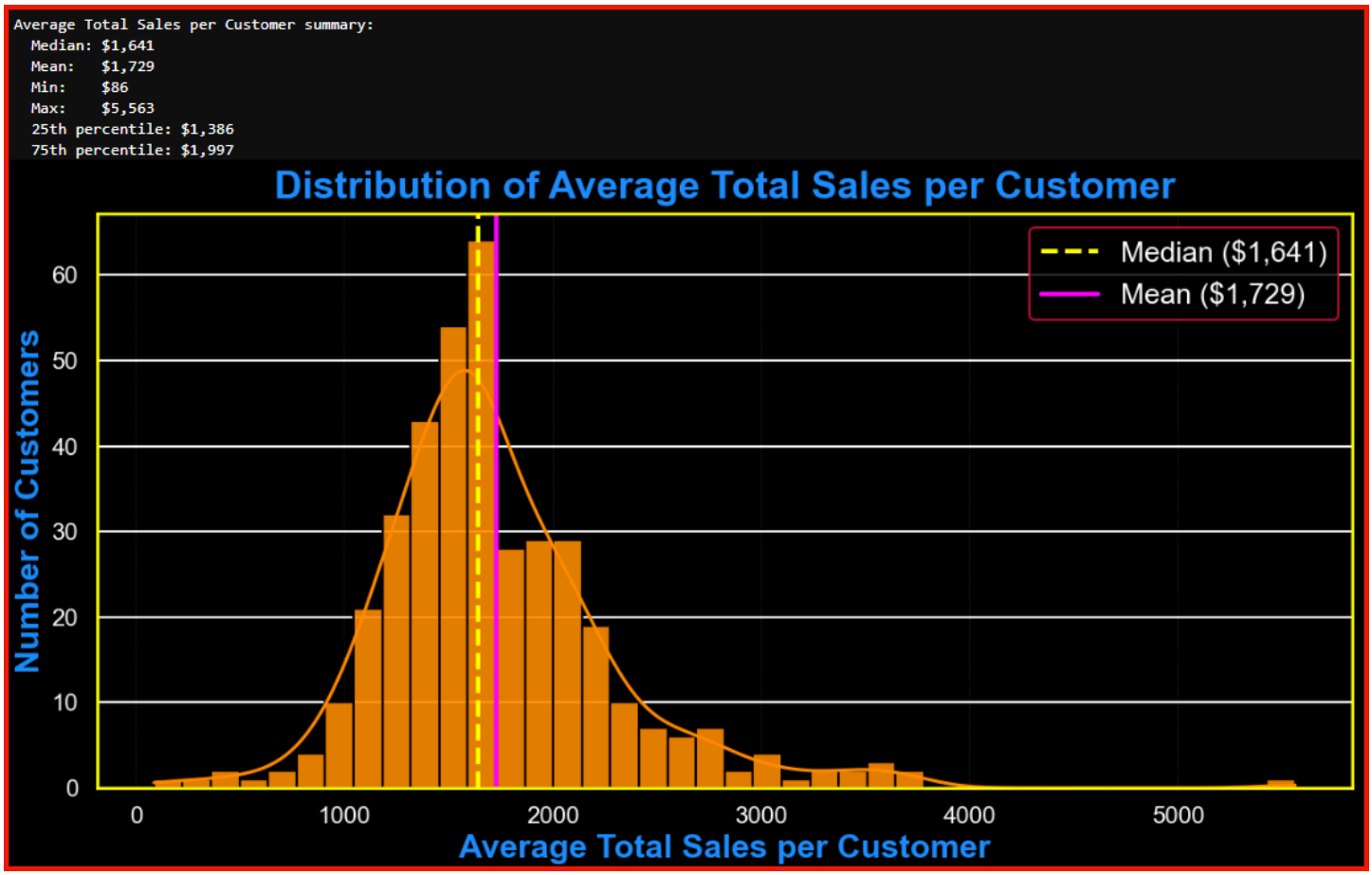
The chart above shows how each customer's average total sales (labor plus parts) is distributed across the business. Most customers generate between $1,400 and $2,000 per invoice on average, with a median value of $1,641 and a mean of $1,729. This tight clustering means that for the majority of accounts, sales per visit are relatively predictable, supporting stable revenue forecasts.

A closer look shows that while the average is stable, there are a few customers at both extremes: a handful bring in much larger average sales (up to $5,563 per invoice), while a small number spend well below $1,000 per visit. This spread highlights the value of not only retaining high-average accounts but also looking for ways to increase the sales per visit for lower-performing customers.

For management, the practical takeaway is clear. Most customer relationships are delivering consistent value, but the biggest opportunities for sales growth come from:

- Lifting the average sales of customers currently below the median
- Identifying and emulating the factors that drive the top-performing accounts

With most customers falling within a fairly narrow sales band, even small improvements in the average can translate into significant revenue gains at scale. The chart also signals that outlier customers, either very high or low, should be reviewed for best practices or potential risks. This information will guide segmentation, targeted sales campaigns, and customer development strategies.



This customer-level sales view sets the stage for deeper exploration. Next, we examine key performance drivers such as labor gross margin, efficiency, and service type to see which factors most strongly influence sales outcomes across our customer base.

3.5: Boxplot: Average Labor Gross Margin % per Customer

This boxplot displays the distribution of average labor gross margin percent (GM%) across all customers in the dataset. For each customer, the GM% is calculated as their average labor margin across all invoices, regardless of job type. In this dataset, that means the boxplot includes all service jobs with billed labor, but excludes any customers who never had a labor charge.

Key findings:

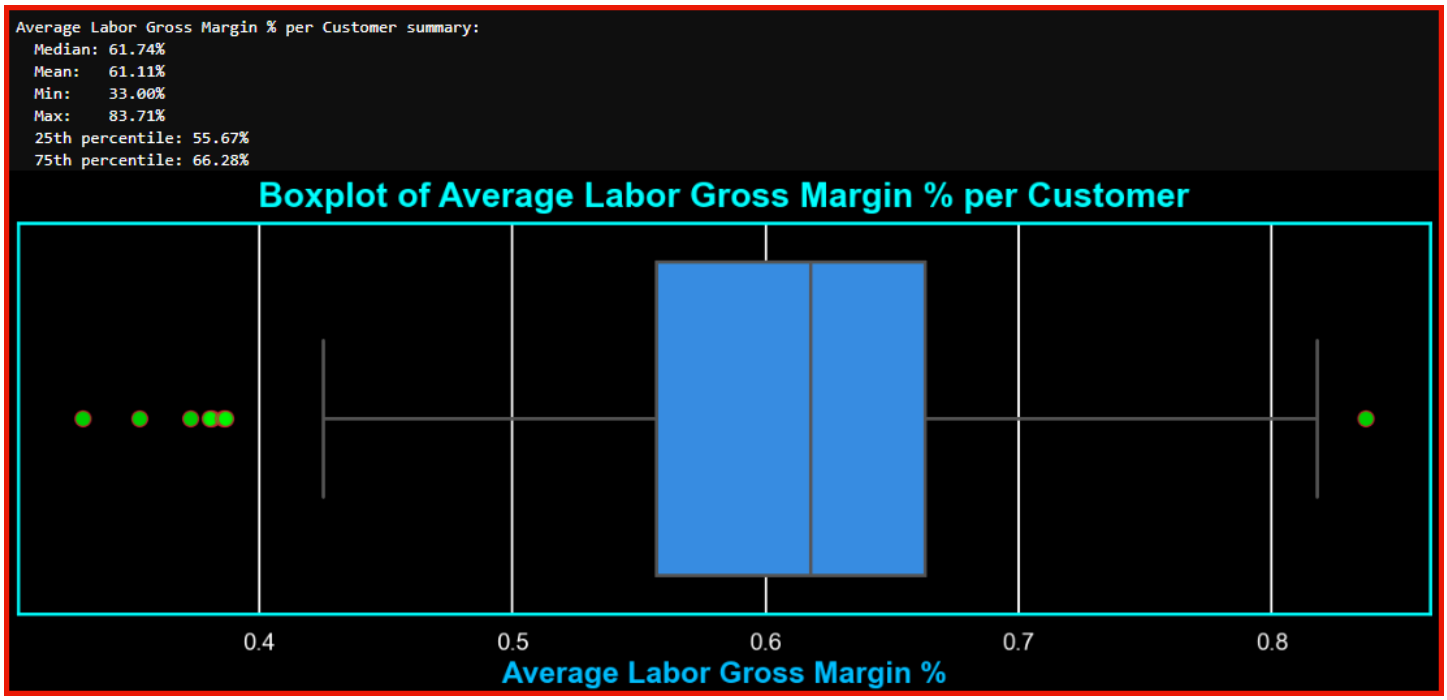
- **Median labor GM% per customer:** 61.74%
- **Mean labor GM% per customer:** 61.11%
- **Range:** From a low of 33% to a high of 84%
- **Most customers** fall between 56% and 66% GM (25th to 75th percentile)

Business implication:

- These margins are typical of a healthy dealership service operation, where most labor jobs are performed with consistent profitability.
- The handful of customers with unusually low or high labor margins likely represent special cases (e.g., warranty work, major discounts, or high-value specialized jobs).
- There are no major outliers on the low end, confirming that the business avoids unprofitable labor work.

- The narrow spread between the 25th and 75th percentiles suggests relatively little variation between customers, which makes sense given standard labor pricing policies.

While this plot confirms consistency in labor profitability, it is less useful for business decision-making than previous plots (e.g., average sales per customer). Most customers cluster tightly in the same GM% range, and the plot doesn't reveal meaningful segments or trends to act on. For this reason, labor gross margin was included as a feature for completeness, but is not expected to be a strong predictor of customer value.



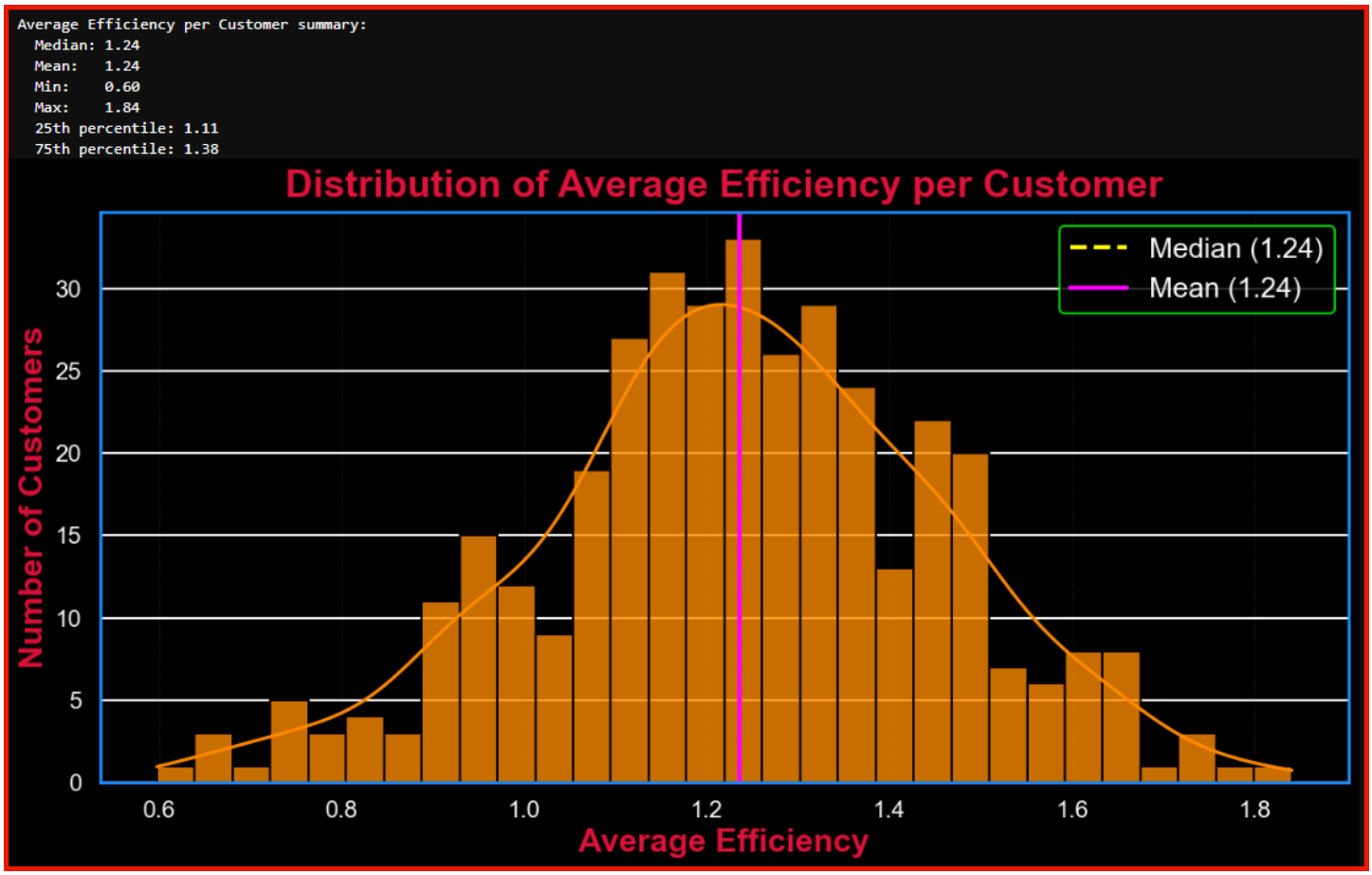
3.6: Distribution of Average Efficiency per Customer

This histogram shows the distribution of average efficiency across all customers in the dataset. Both the mean and median efficiency are 1.24, with the majority of customers falling between 1.11 and 1.38. The minimum observed is 0.60, and the maximum is 1.84. These results suggest that most customers are billed for slightly more hours than they actually work, which is typical in dealership operations due to standard labor practices, billing policies, and rounding.

However, it's important to note that in real dealership data, it is highly unlikely for the mean and median efficiency to match exactly. In a real-world environment, you would expect greater variation and more outliers, reflecting differences in technician performance, customer mix, and possible data entry errors. The close match between mean and median in this analysis is a limitation of using programmatically generated (simulated) data, which tends to produce more regular and "clean" distributions. When this script is applied to actual customer data, it will reveal a more realistic and varied efficiency distribution.

Business implication:

- Most customers are operating near or slightly above target efficiency.
- The lack of extreme outliers indicates consistent billing and operational practices.
- Future analysis with real data will provide more actionable insights and flag potential opportunities for coaching, process improvement, or audit.



3.7: Correlation Matrix

I created a correlation matrix to check how the key customer-level features move together. This is a standard EDA step before modeling, but not every correlation is business-relevant. Here's what stands out:

- **AvgTotalSalesPerInvoice and AvgLaborGM**: There is a moderate positive correlation (0.40). This suggests that customers with higher average sales per invoice also tend to have higher labor gross margins. In practice, this means that more profitable work often goes hand-in-hand with larger sales, which is good for both revenue and margin goals.
- **AvgTotalSalesPerInvoice and AvgPartsGM**: Weakly positive (0.18). While the relationship is not strong, higher parts margins are still somewhat associated with better overall sales per invoice.
- **AvgTotalSalesPerInvoice and TotalInvoices**: Almost zero correlation. This is unexpected. In real data, we might expect that more frequent customers would spend more per visit, but in this dataset, invoice frequency does not predict invoice size.
- **AvgTotalSalesPerInvoice and AvgEfficiency**: Slightly negative (-0.07). This means efficiency (hours billed vs. hours worked) has almost no relationship to average customer sales in this sample.

The main business takeaway from this correlation analysis is that gross margin, especially on labor, is a better indicator of higher-value customers than invoice frequency or efficiency. Efficiency and invoice count are important operational KPIs but do not, in this dataset, predict higher sales or margins on their own.

This analysis also helps you avoid multicollinearity when building the regression model, as none of the variables are strongly redundant. You can keep all features for now, but the weak relationships suggest the model may need additional data or more powerful features to predict sales with high accuracy.

Labor margin and parts margin are worth close attention for growing profitable customer segments.
Frequency of visits and efficiency, at least in this data, have little direct effect on average sales per customer.
These results may differ with real dealership data but provide a useful baseline for evaluating customer value.

Correlation Matrix Summary:

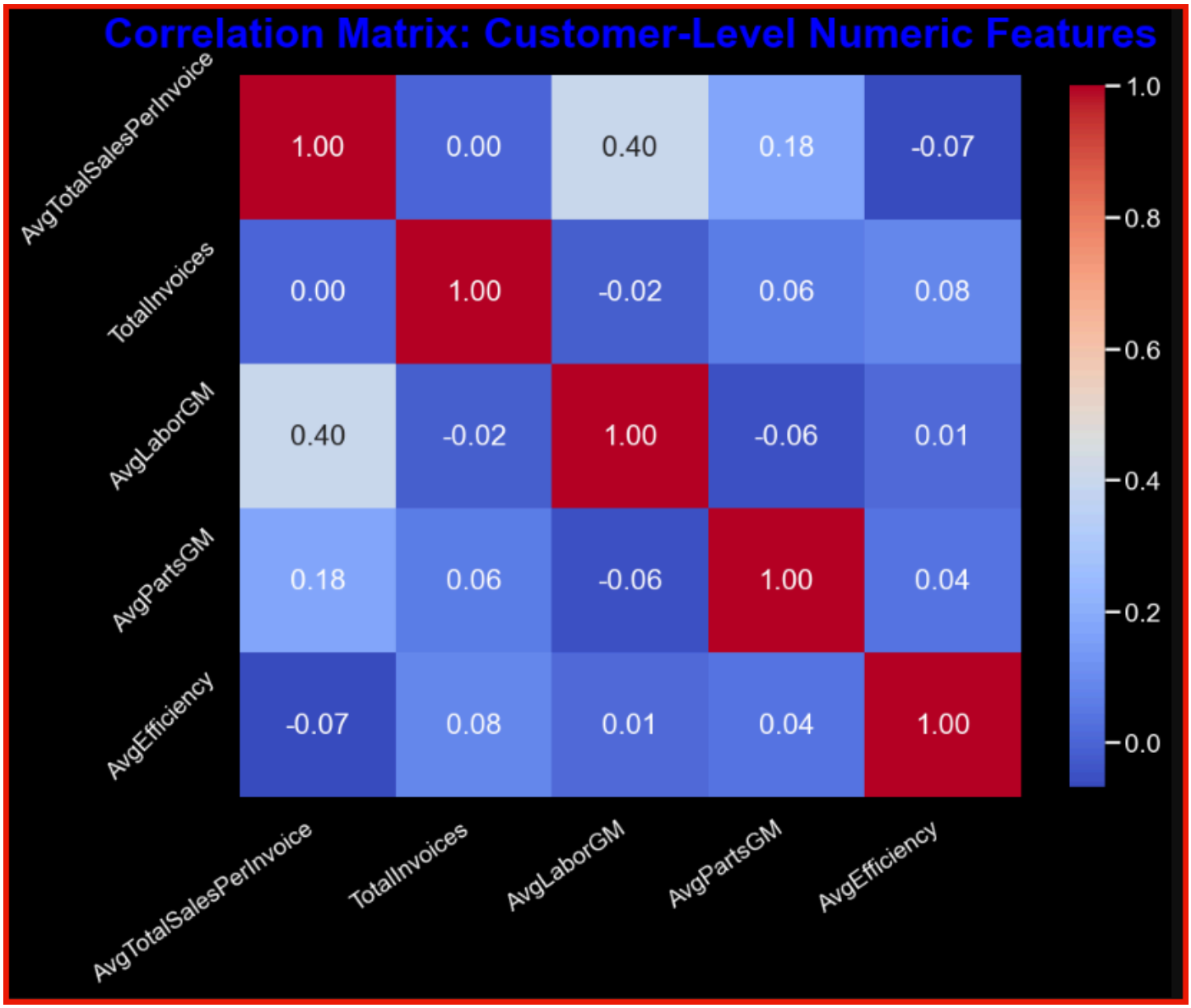
AvgTotalSalesPerInvoice vs. TotalInvoices : 0.00
AvgTotalSalesPerInvoice vs. AvgLaborGM : 0.40
AvgTotalSalesPerInvoice vs. AvgPartsGM : 0.18
AvgTotalSalesPerInvoice vs. AvgEfficiency : -0.07

TotalInvoices vs. AvgTotalSalesPerInvoice: 0.00
TotalInvoices vs. AvgLaborGM : -0.02
TotalInvoices vs. AvgPartsGM : 0.06
TotalInvoices vs. AvgEfficiency : 0.08

AvgLaborGM vs. AvgTotalSalesPerInvoice: 0.40
AvgLaborGM vs. TotalInvoices : -0.02
AvgLaborGM vs. AvgPartsGM : -0.06
AvgLaborGM vs. AvgEfficiency : 0.01

AvgPartsGM vs. AvgTotalSalesPerInvoice: 0.18
AvgPartsGM vs. TotalInvoices : 0.06
AvgPartsGM vs. AvgLaborGM : -0.06
AvgPartsGM vs. AvgEfficiency : 0.04

AvgEfficiency vs. AvgTotalSalesPerInvoice: -0.07
AvgEfficiency vs. TotalInvoices : 0.08
AvgEfficiency vs. AvgLaborGM : 0.01
AvgEfficiency vs. AvgPartsGM : 0.04



The next EDA step drills down into specific feature relationships using scatter plots, allowing us to see if any nonlinear or segmented patterns might be missed by simple correlation analysis.

3.8: Total Invoices vs. Average Total Sales per Customer

This scatter plot maps each customer's total number of invoices against their average total sales per invoice. Business leaders can instantly spot which accounts drive volume, which deliver high sales per job, and which do both. The top three customers by invoice count (Paramount Transfer, Acme Inc, and Paramount Logistics) appear on the far right. Each of these customers generates a large number of invoices, but their average sales per job remain near the company median. Signaling consistent, repeat business at typical sales values.

In contrast, customers with the highest average sales per invoice (like Pioneer Associates and Keystone Tucking) are clustered at the top left. Most of these are small or one-off accounts with just a handful of jobs, often tied to unique or high-value projects. Only one high-average customer, Sterling Systems, also appears with a large invoice count, an outlier who deserves special attention.

Business Implications:

- Most of the company's stable revenue comes from high-invoice customers who buy at predictable, moderate values.
- High-average, low-volume customers bring in big jobs but do not provide consistent revenue.
- Efforts to grow revenue should focus on two strategies: deepen relationships with top invoice-count customers and identify what makes high-average customers unique.
- Special attention should be paid to customers like Sterling Systems, who combine both scale and high sales per job, these are rare and valuable accounts.

The logarithmic scale version of this plot was tested, but it did not add business value; it compressed the data without revealing new patterns. Visualizing customer names alongside the top points on the plot helps managers link the data to real accounts, making the insights more actionable.

Total Invoices per Customer summary:

- Median: 28
- Mean: 117.6
- Min: 1
- Max: 2008
- 25th percentile: 13
- 75th percentile: 68

Average Total Sales per Customer summary:

- Median: $1,641
- Mean: $1,729
- Min: $86
- Max: $5,563
- 25th percentile: $1,386
- 75th percentile: $1,997

Top 10 Customers by Total Invoices:

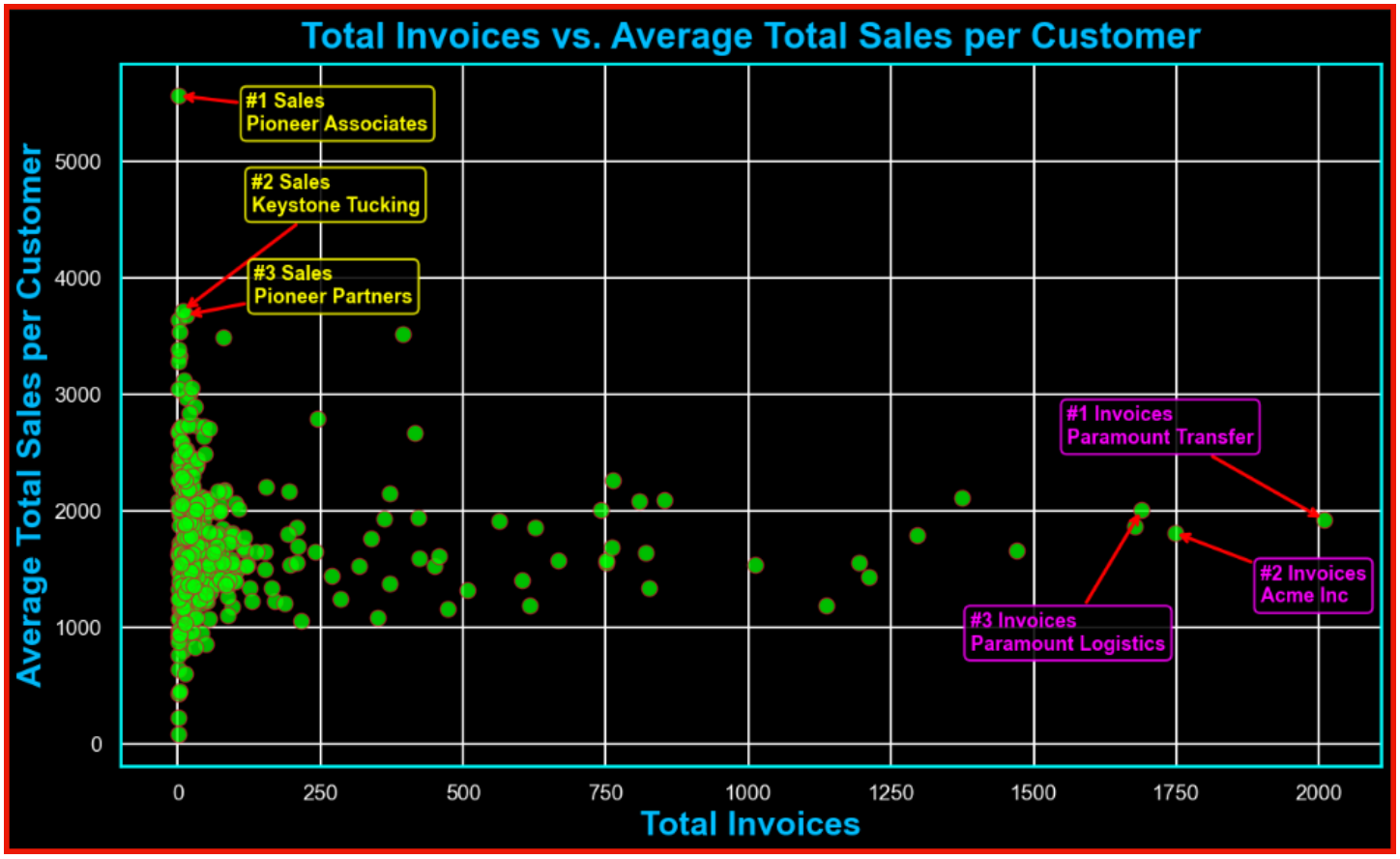
Paramount Transfer | Invoices: 2008 | Avg Total Sales: $1,927
Acme Inc | Invoices: 1747 | Avg Total Sales: $1,814
Paramount Logistics | Invoices: 1689 | Avg Total Sales: $2,007
Liberty Industries | Invoices: 1677 | Avg Total Sales: $1,867
Goldstar Transfer | Invoices: 1470 | Avg Total Sales: $1,662
Stonegate Partners | Invoices: 1375 | Avg Total Sales: $2,113
Unity Resources | Invoices: 1295 | Avg Total Sales: $1,797
Evergreen Associates | Invoices: 1212 | Avg Total Sales: $1,432
Aurora Inc | Invoices: 1193 | Avg Total Sales: $1,558
Evergreen LLC | Invoices: 1136 | Avg Total Sales: $1,186

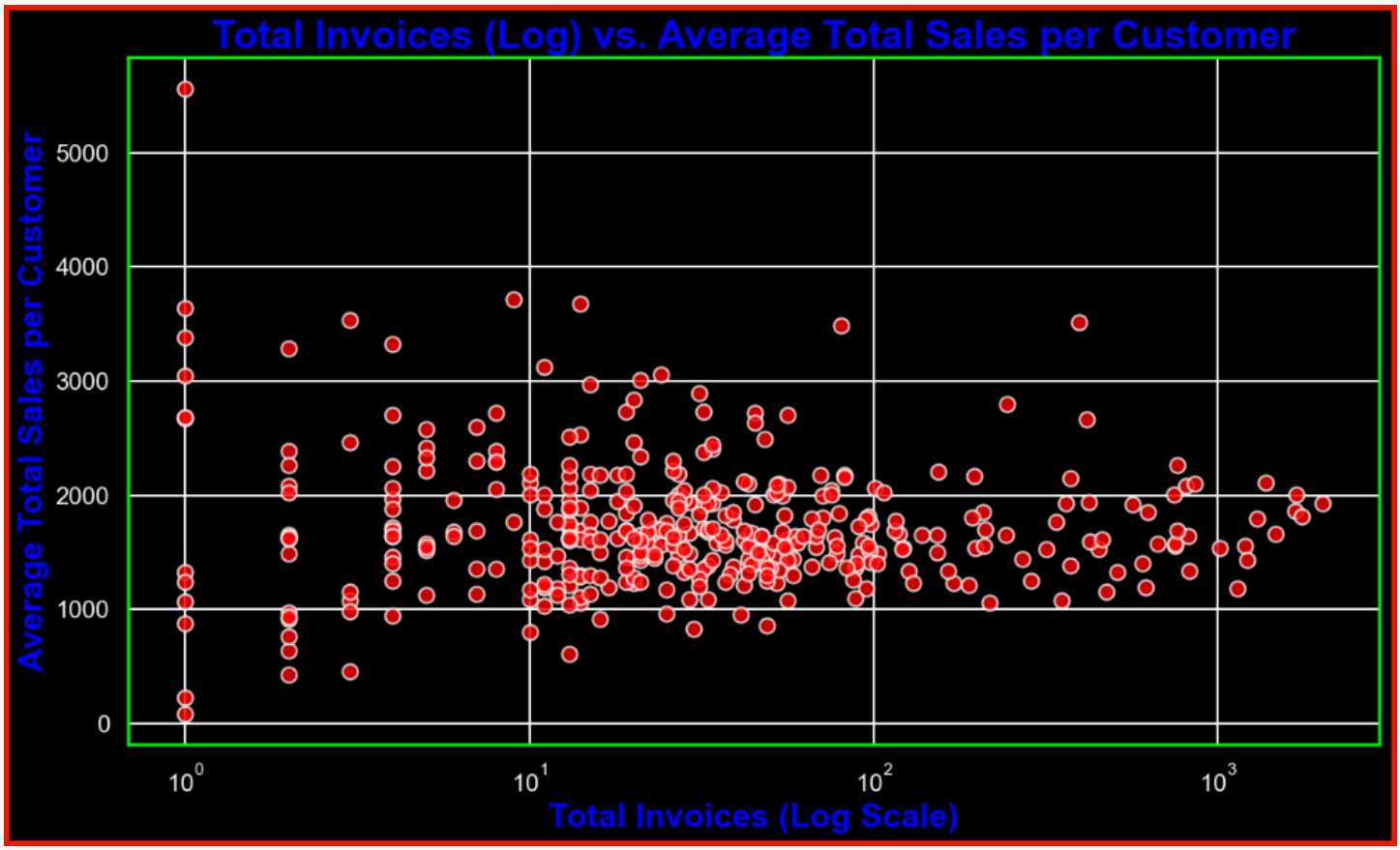
Top 10 Customers by Average Total Sales per Invoice:

Pioneer Associates | Avg: $5,563 | Invoices: 1
Keystone Tucking | Avg: $3,718 | Invoices: 9
Pioneer Partners | Avg: $3,681 | Invoices: 14

Pioneer Truck Lines | Avg: $3,637 | Invoices: 1
Evergreen Transfer | Avg: $3,539 | Invoices: 3
Sterling Systems | Avg: $3,520 | Invoices: 395
Synergy Solutions | Avg: $3,492 | Invoices: 80
Sterling Corp | Avg: $3,383 | Invoices: 1
Evergreen Transfer | Avg: $3,329 | Invoices: 4
Keystone Global | Avg: $3,287 | Invoices: 2

- Grow share with high-invoice customers by exploring cross-sell, up-sell, or loyalty programs.
- Investigate high-average customers to identify successful project types or pricing models that could be scaled.
- Use these visualizations in management reviews to connect data directly to account strategy.





3.9: Total Service Invoices vs. Average Labor Sales per Customer

This chart shows the relationship between total service invoices and average labor sales per service customer. Paramount Transfer stands out as the clear leader in service volume, with over 1,100 service invoices. Their average labor sales per invoice is just above the median for all service customers, indicating they consistently bring in steady work at a healthy, but not excessive, value per job. Paramount Logistics and Liberty Industries also show high invoice counts with respectable average labor sales.

On the other hand, several customers like Pioneer Partners and Synergy Solutions have much higher average labor sales per invoice, but only a handful of total jobs. These outliers may represent specialized or project-based work, rather than consistent repeat business.

Business Implication:
High-volume service customers such as Paramount Transfer are the foundation of our service revenue. Maintaining strong relationships, offering dedicated support, and identifying ways to grow their volume or increase their average sale could yield significant gains. While big-ticket, low-frequency customers can be profitable, the stability and predictability from high-volume accounts make them more valuable over the long run. Management attention should prioritize retention and upsell strategies for top service customers, rather than chasing sporadic high-dollar sales.

Average Labor Sales per Service Customer summary:

- Median: $1,920
- Mean: $2,057
- Min: $729
- Max: $5,487

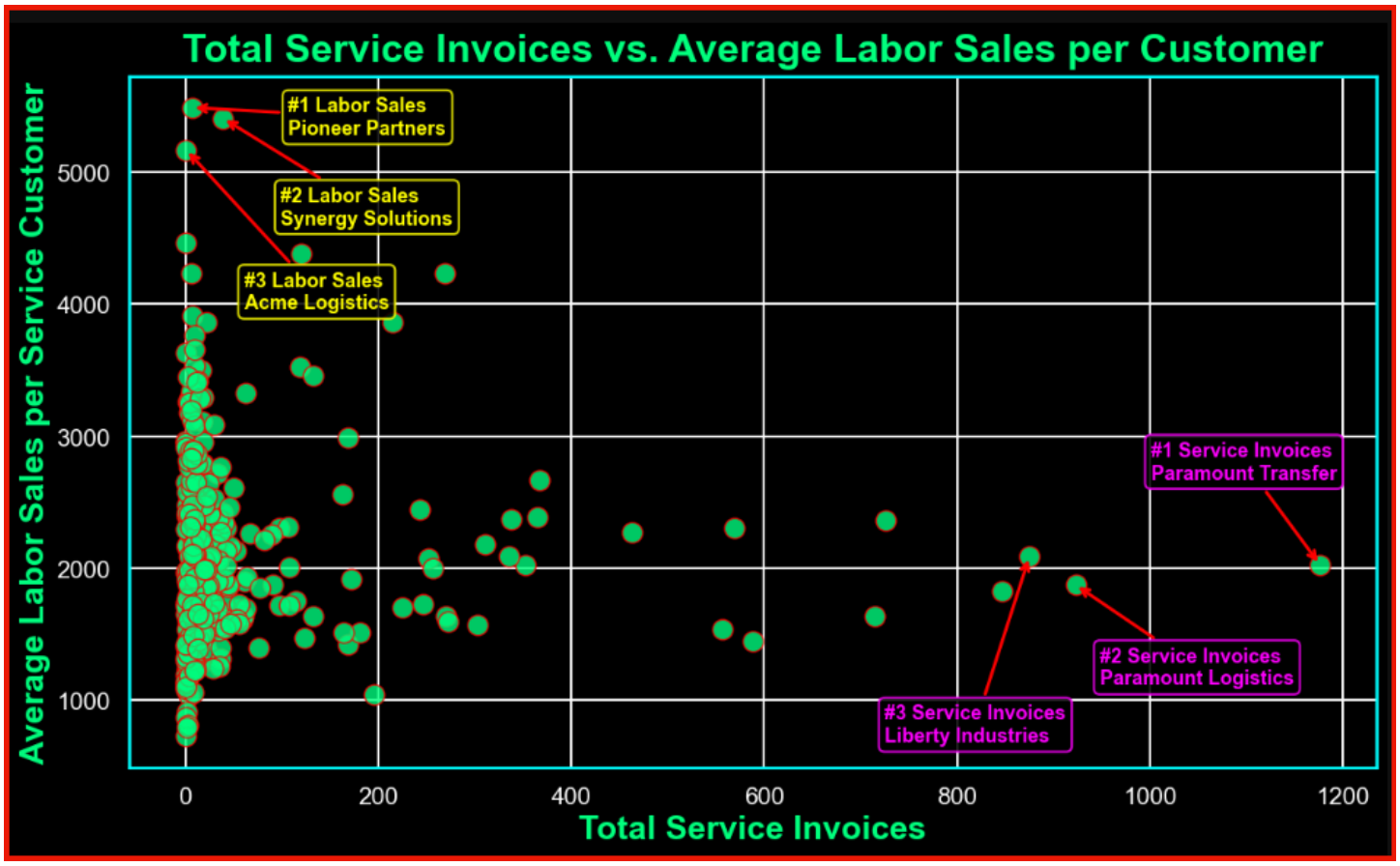
- 25th percentile: $1,621
- 75th percentile: $2,341

Top 10 Customers by Total Service Invoices:

19. Paramount Transfer | Service Invoices: 1176 | Avg Labor Sales: $2,028
31. Paramount Logistics | Service Invoices: 923 | Avg Labor Sales: $1,872
82. Liberty Industries | Service Invoices: 875 | Avg Labor Sales: $2,088
118. Acme Inc | Service Invoices: 847 | Avg Labor Sales: $1,830
146. Stonegate Partners | Service Invoices: 726 | Avg Labor Sales: $2,366
137. Goldstar Transfer | Service Invoices: 715 | Avg Labor Sales: $1,635
32. Aurora Inc | Service Invoices: 588 | Avg Labor Sales: $1,444
301. Unity Resources | Service Invoices: 569 | Avg Labor Sales: $2,306
295. Evergreen Associates | Service Invoices: 557 | Avg Labor Sales: $1,538
353. Acme Systems | Service Invoices: 463 | Avg Labor Sales: $2,273

Top 10 Customers by Avg Labor Sales per Service Invoice:

303. Pioneer Partners | Avg Labor Sales: $5,487 | Service Invoices: 7
144. Synergy Solutions | Avg Labor Sales: $5,406 | Service Invoices: 39
95. Acme Logistics | Avg Labor Sales: $5,165 | Service Invoices: 1
41. Pioneer Associates | Avg Labor Sales: $4,464 | Service Invoices: 1
26. Stonegate Truck Lines | Avg Labor Sales: $4,382 | Service Invoices: 120
317. Keystone Tucking | Avg Labor Sales: $4,235 | Service Invoices: 6
297. Sterling Systems | Avg Labor Sales: $4,234 | Service Invoices: 269
36. Paramount Global | Avg Labor Sales: $3,911 | Service Invoices: 7
240. Synergy Transfer | Avg Labor Sales: $3,864 | Service Invoices: 215
200. Pioneer Partners | Avg Labor Sales: $3,862 | Service Invoices: 22



3.10: Total Invoices vs. Average Parts Sales per Customer

This plot focuses on all invoices (parts and service) and average parts sales per customer. Paramount Transfer again leads in total invoice volume but falls below the median for parts sales per invoice. In contrast, Acme Inc and Paramount Logistics (#2 and #3 in invoice count) both post average parts sales per invoice above the median, making them especially valuable to the business.

A small number of customers, such as Goldstar Solutions and Unity Inc, post the highest average parts sales per invoice, but with very few transactions. These may be bulk orders or special projects, not regular buyers.

Business Implication:
Customers with both high invoice counts and above-average parts sales (like Acme Inc and Paramount Logistics) represent the ideal profile, repeat business with strong revenue per transaction. These accounts deserve focused account management, potential loyalty incentives, and frequent check-ins to ensure retention.
For Paramount Transfer, the leading invoice customer, the low average parts sale per invoice suggests room for operational improvement. Working with them to consolidate orders could increase average sales per invoice, reduce transactional overhead, and improve efficiency for both companies.

For those customers with high average sales but low invoice counts, it may be worth exploring why they aren't purchasing more frequently and if there are barriers preventing them from becoming repeat buyers.

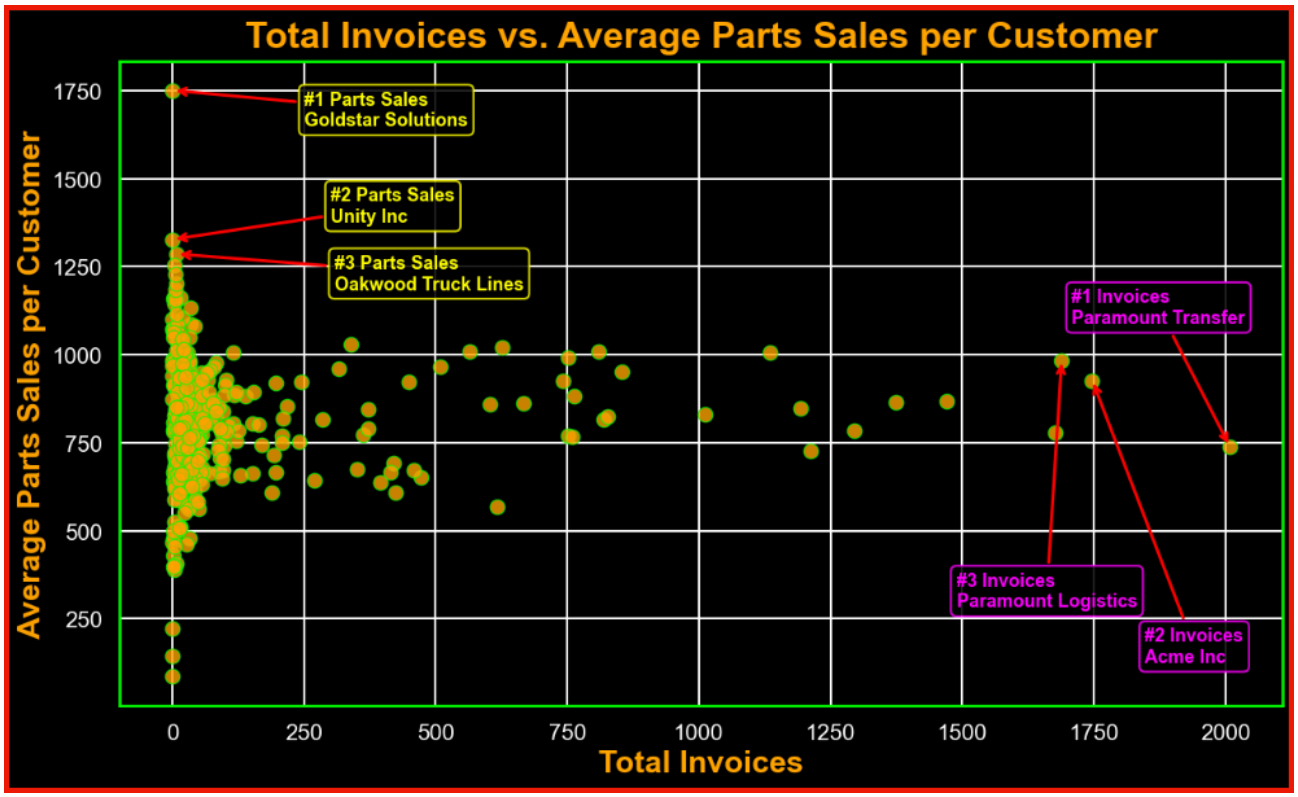
Average Parts Sales per Customer summary:

- Median: $802

- Mean: $804
- Min: $86
- Max: $1,750
- 25th percentile: $697
- 75th percentile: $893

Top 10 Customers by Total Invoices:
21. Paramount Transfer | Invoices: 2008 | Avg Parts Sales: $739
121. Acme Inc | Invoices: 1747 | Avg Parts Sales: $927
34. Paramount Logistics | Invoices: 1689 | Avg Parts Sales: $984
85. Liberty Industries | Invoices: 1677 | Avg Parts Sales: $777
142. Goldstar Transfer | Invoices: 1470 | Avg Parts Sales: $867
151. Stonegate Partners | Invoices: 1375 | Avg Parts Sales: $864
311. Unity Resources | Invoices: 1295 | Avg Parts Sales: $784
305. Evergreen Associates | Invoices: 1212 | Avg Parts Sales: $725
35. Aurora Inc | Invoices: 1193 | Avg Parts Sales: $846
10. Evergreen LLC | Invoices: 1136 | Avg Parts Sales: $1,005

Top 10 Customers by Avg Parts Sales per Invoice:
275. Goldstar Solutions | Avg Parts Sales: $1,750 | Invoices: 1
159. Unity Inc | Avg Parts Sales: $1,327 | Invoices: 1
82. Oakwood Truck Lines | Avg Parts Sales: $1,286 | Invoices: 8
205. Sunset Solutions | Avg Parts Sales: $1,254 | Invoices: 3
381. Evergreen Tucking | Avg Parts Sales: $1,228 | Invoices: 5
283. Liberty Inc | Avg Parts Sales: $1,203 | Invoices: 8
207. Summit Tucking | Avg Parts Sales: $1,183 | Invoices: 5
37. Evergreen Transfer | Avg Parts Sales: $1,164 | Invoices: 4
141. Goldstar Global | Avg Parts Sales: $1,161 | Invoices: 15
16. Stonegate Group | Avg Parts Sales: $1,158 | Invoices: 2



3.12: Average Efficiency vs. Average Total Sales per Customer

This scatter plot shows each customer's average efficiency (hours billed divided by hours worked) alongside their average total sales per invoice. Two sets of outliers are highlighted: the top three customers by efficiency and the top three by sales.

Most customers fall within a moderate efficiency range (median: 1.24), but some stand out with efficiency values above 1.7. The top 10 most efficient customers all show efficiency above 1.6, but only a few of them also appear among the top 10 by sales per invoice. In other words, being highly efficient does not always translate to the highest revenue per transaction.

Similarly, the top 10 customers by average total sales per invoice show a wide range of efficiency, from below 1.0 to well above 1.5. Several customers, such as Pioneer Associates and Sterling Systems, stand out for high average sales per invoice, but are not among the most efficient. A few, like Pioneer Partners and Evergreen Transfer, combine both high efficiency and strong average sales.

This analysis is valuable for business planning:

- **Actionable insight:** Customers who combine high sales per invoice with high efficiency are especially valuable. Targeting these accounts for retention and growth could have a significant impact.
- **Opportunity:** Some very efficient customers may not be generating the highest sales. These could be cross-sell or up-sell targets.
- **Limitation:** The dataset's efficiency values are tightly grouped around the mean, likely a result of the synthetic data generation process. In real-world data, greater spread and more dramatic outliers would be expected.

By segmenting customers in this way, management can set goals not only for total sales, but also for operational efficiency, focusing on accounts that drive both revenue and process excellence.

Total Invoices per Customer summary:

- Median: 28
- Mean: 117.6
- Min: 1
- Max: 2008
- 25th percentile: 13
- 75th percentile: 68

Average Total Sales per Customer summary:

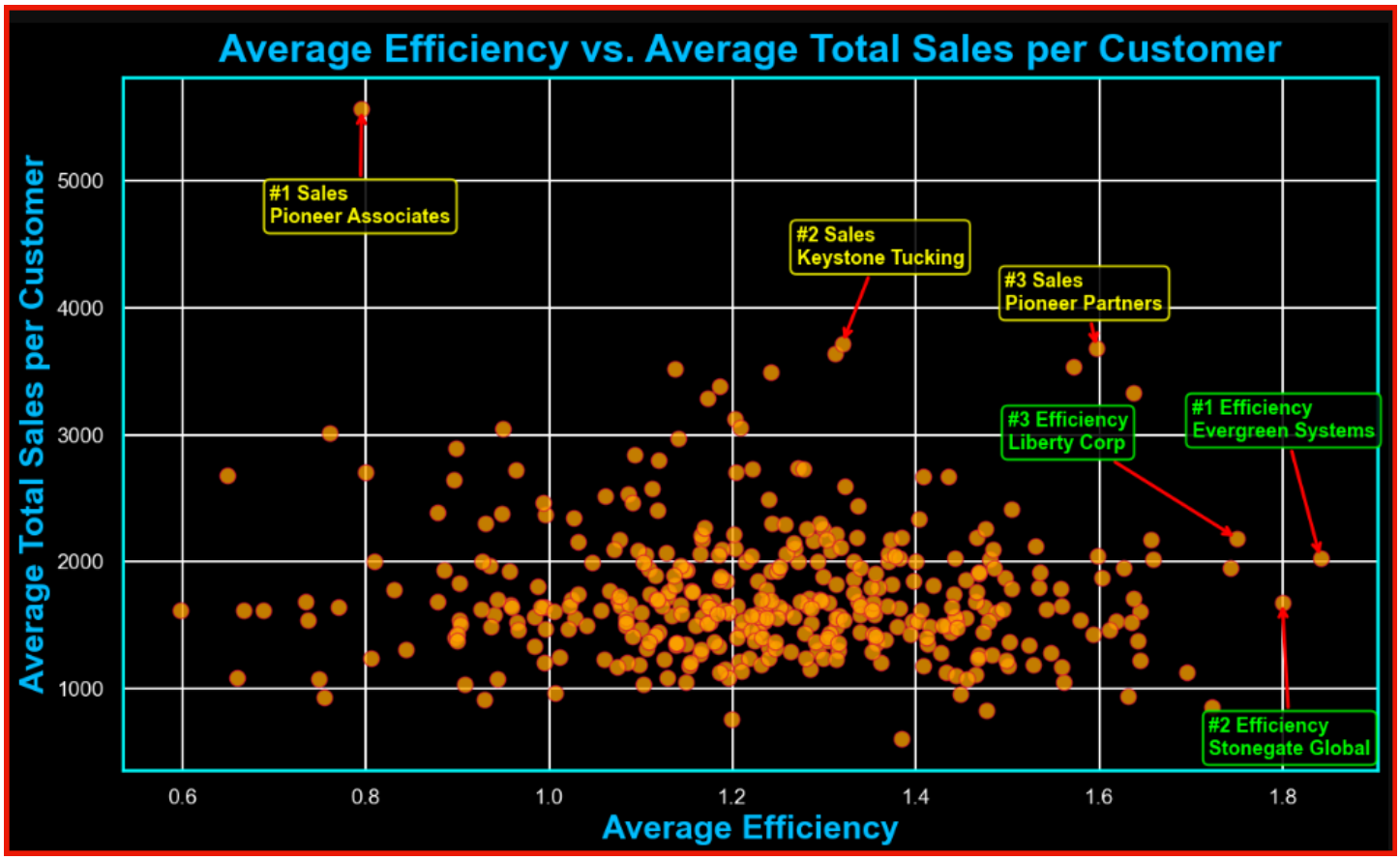
- Median: $1,641
- Mean: $1,729
- Min: $86
- Max: $5,563
- 25th percentile: $1,386
- 75th percentile: $1,997

Top 10 Customers by Total Invoices:

21. Paramount Transfer | Invoices: 2008 | Avg Total Sales: $1,927
121. Acme Inc | Invoices: 1747 | Avg Total Sales: $1,814
34. Paramount Logistics | Invoices: 1689 | Avg Total Sales: $2,007
85. Liberty Industries | Invoices: 1677 | Avg Total Sales: $1,867
142. Goldstar Transfer | Invoices: 1470 | Avg Total Sales: $1,662
151. Stonegate Partners | Invoices: 1375 | Avg Total Sales: $2,113
311. Unity Resources | Invoices: 1295 | Avg Total Sales: $1,797
305. Evergreen Associates | Invoices: 1212 | Avg Total Sales: $1,432
35. Aurora Inc | Invoices: 1193 | Avg Total Sales: $1,558
10. Evergreen LLC | Invoices: 1136 | Avg Total Sales: $1,186

Top 10 Customers by Average Total Sales per Invoice:

44. Pioneer Associates | Avg Total Sales: $5,563 | Invoices: 1
327. Keystone Tucking | Avg Total Sales: $3,718 | Invoices: 9
313. Pioneer Partners | Avg Total Sales: $3,681 | Invoices: 14
45. Pioneer Truck Lines | Avg Total Sales: $3,637 | Invoices: 1
221. Evergreen Transfer | Avg Total Sales: $3,539 | Invoices: 3
307. Sterling Systems | Avg Total Sales: $3,520 | Invoices: 395
149. Synergy Solutions | Avg Total Sales: $3,492 | Invoices: 80
192. Sterling Corp | Avg Total Sales: $3,383 | Invoices: 1
37. Evergreen Transfer | Avg Total Sales: $3,329 | Invoices: 4
65. Keystone Global | Avg Total Sales: $3,287 | Invoices: 2



3.13: Average Total Sales per Customer by Most Common ROtype

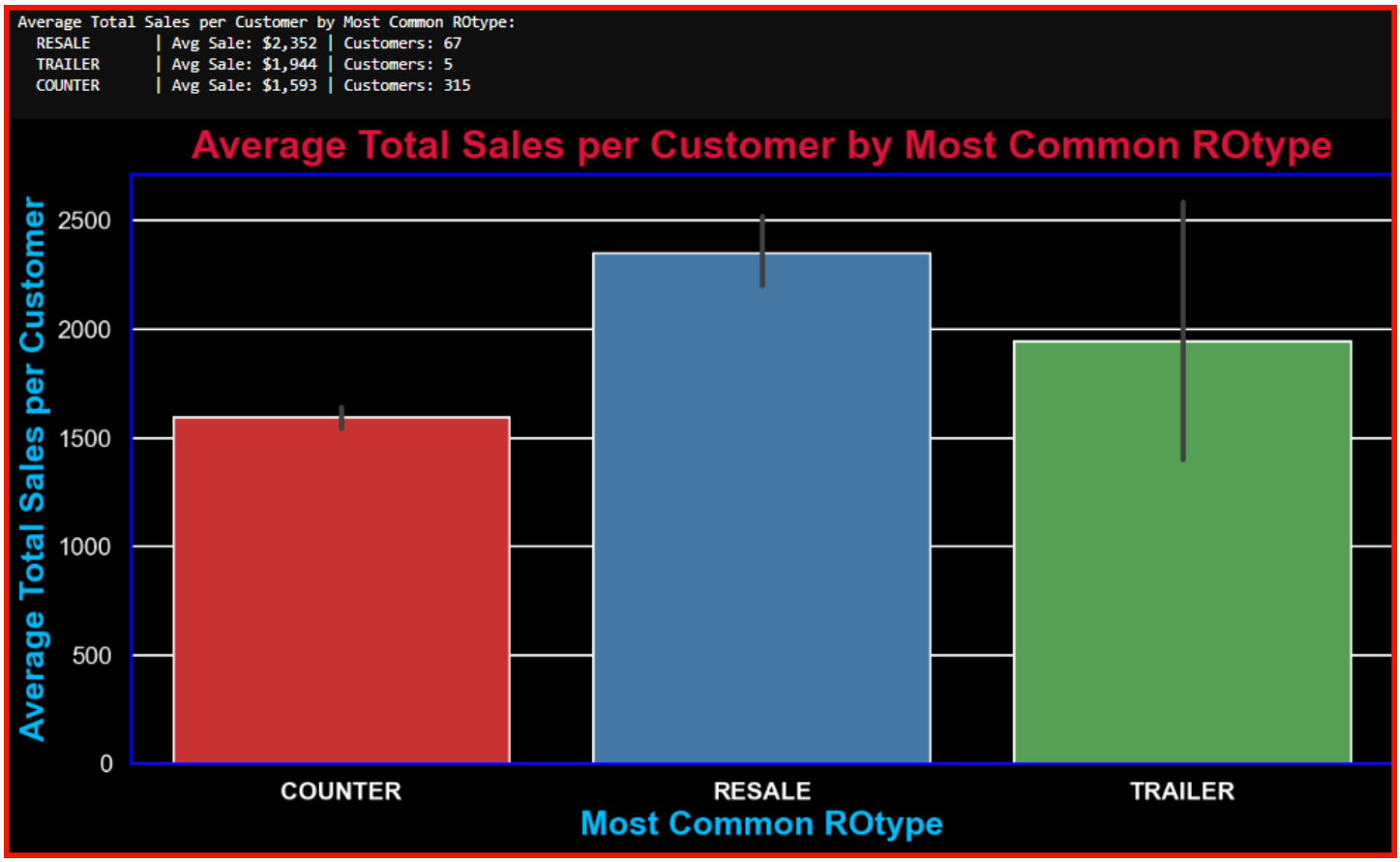
This chart breaks down customers by the most common type of repair order (ROtype) associated with their business: **COUNTER, RESALE, or TRAILER**. The data show a clear pattern:

- **RESALE** customers generate the highest average total sales per customer ($2,352), despite being a smaller group (67 customers).
- **TRAILER** is a much smaller segment (5 customers), but their average sales ($1,944) are noticeably higher than the COUNTER group.
- **COUNTER** customers make up the largest group (315 customers) but have the lowest average sales per customer ($1,593).

Business Interpretation:

- The higher average for RESALE customers suggests that accounts with frequent resale work (which includes both labor and parts sales) drive more value for the dealership.
- COUNTER accounts, while numerous, tend to have lower-dollar transactions, likely dominated by parts-only sales and smaller jobs.
- TRAILER jobs also show strong average sales, but the small number of customers in this group means any business decisions based on this segment should be cautious.

The dealership may want to target more customers for RESALE work or create strategies to convert COUNTER customers into higher-value accounts. Marketing, sales training, or customer incentives could be focused on increasing RESALE-type work to lift overall average sales per customer.



3.14: Statistical Comparison: RESALE vs. COUNTER Customers

The chart above shows that customers whose primary business is RESALE work have a higher average total sales per customer than those focused on COUNTER transactions. To confirm whether this observed difference is statistically meaningful, and not just due to random chance, we performed a two-sample t-test comparing the average total sales per customer for the RESALE and COUNTER groups.

For this statistical comparison, I limited the analysis to customers whose most common ROtype is either RESALE or COUNTER. These two categories represent the vast majority of customers in the dataset. RESALE customers typically account for full-service jobs involving both labor and parts, while COUNTER customers primarily purchase parts only. The TRAILER category, by contrast, included only a handful of customers, making statistical analysis unreliable due to the small sample size. By focusing on RESALE and COUNTER customers, the results are both meaningful and robust for this business context.

If future datasets include more TRAILER or TRUCK customers, similar analyses could be repeated for those groups.

These are the results:

- T-statistic: 11.407577033806557
- P-value: 4.0199041135806196e-26

The extremely low p-value (far below 0.05) shows that the difference in average total sales per customer between RESALE and COUNTER groups is highly statistically significant. In practical terms, this means that RESALE-focused customers reliably bring in more revenue per account than COUNTER-focused customers, and this difference is very unlikely to be due to random variation in the data. For management, this result

supports prioritizing RESALE business development and deepening relationships with these higher-value accounts.

3.15: Summary and Interpretation

The exploratory data analysis (EDA) revealed clear patterns in customer performance. The two strongest indicators of higher average sales per customer are the number of invoices (customer activity level) and the most common job type (RESALE, COUNTER, etc.). Customers who bring in more jobs and who have a higher proportion of full-service work tend to generate the most revenue.

Average efficiency per customer shows a small but positive relationship with sales, suggesting that more efficient accounts may do slightly higher value work on average.

Gross margin percentages for labor and parts, while important for profit, do not have a strong direct impact on average customer sales in this dataset. These fields are more useful for downstream profitability analysis than for predicting customer sales totals.

These insights will guide our selection of input features for regression modeling. We now have realistic expectations for prediction accuracy, and have identified which customer attributes are most useful for targeting high-value accounts and segmenting the customer base. The EDA results also suggest opportunities for business action. Such as focusing retention efforts on active, full-service customers, or investigating ways to grow invoice volume with under-performing but high-potential accounts.

Section 4: Model Development and Evaluation

4.1: Prepare Data for Modeling

The final feature matrix contained 387 customer records and 61 engineered features, including invoice counts, sales averages, labor/parts gross margins, efficiency, and one-hot-encoded categorical variables for repair type, department, and location.

4.2: Train/Test Split

To ensure that model evaluation is unbiased and generalized, the customer dataset was split into training and test sets using an 80/20 split. The training set (309 customers, ~80% of the data) is used to fit the regression models, while the test set (78 customers, ~20%) is reserved for out-of-sample validation. This approach allows us to measure the model's ability to predict average customer sales on new, unseen customers.

Train/Test Split Results:

- X_train shape: (309, 24)
- X_test shape: (78, 24)
- y_train shape: (309,)
- y_test shape: (78,)
- Training set percent: 79.8%
- Test set percent: 20.2%

With the data split into train and test sets, we now fit, evaluate, and interpret multiple regression models to predict customer sales performance.

4.3: Model 1: Linear Regression

Approach:
The first model used was linear regression. This approach creates a prediction for each customer's average total sales per invoice based on customer-level features. The features included invoice counts, average efficiency, gross margins, repair type, department, and location. The data was split so that 80 percent was used to train the model and 20 percent was set aside for testing performance.

Results:

- Number of features used: 17
- Test set size: 78 customers
- R^2 (Test): 0.567384
- MAE (Test): $236.47
- RMSE (Test): $302.15

Interpretation:
The linear regression model explained about 57 percent of the variation in average customer sales. The average prediction error (MAE) was $236.47 per customer, with most predictions within about $300 (RMSE) of the actual values. This means the model is able to capture a good share of the factors that make one customer higher value than another, based on the available data. However, there is still a sizable portion of variability that is not explained by this simple model, which is typical for real-world business data.

Business Implications:
This baseline result suggests that basic customer activity data and operational metrics can predict customer value with moderate accuracy. The model provides a useful starting point for identifying high-value accounts and can help focus sales and retention efforts. More advanced models in the next steps may improve this further by capturing more complex relationships.

4.4: Model 2: Random Forest Regressor

Approach:
The random forest model was chosen as a second, more powerful approach to capture complex, nonlinear relationships between features and customer sales. Random forests can automatically account for interactions and patterns that a linear model might miss.

Results:

- Features used: 17
- Test set size: 78 customers
- R^2 (Test): 0.525416
- MAE (Test): $250.68
- RMSE (Test): $316.47

Interpretation:
The random forest explained about 53 percent of the variation in average customer sales. This is similar to the linear regression result, with a slightly higher error. In this case, the random forest did not substantially outperform the linear model. This could be due to the structure of the data, the amount of noise present, or the fact that the underlying relationships may already be well captured by linear combinations of features.

Business Implications:
Random forest models are useful when feature interactions and nonlinearities are suspected. In this scenario, the added model complexity did not result in a significant accuracy gain. However, the model provides a useful second opinion and can help confirm that the customer-level patterns are stable and not dependent on any one model type.

4.5: Model 3: XGBoost Regressor

Approach:
The XGBoost regressor is a powerful, gradient-boosted tree model that is often used to capture subtle nonlinearities and complex feature interactions. It is widely used for structured business data because of its robustness and flexibility.

Results:

- Features used: 17
- Test set size: 78 customers
- R^2 (Test): 0.426515
- MAE (Test): $269.12
- RMSE (Test): $347.88

Interpretation:
The XGBoost model explained about 43 percent of the variation in average customer sales. While XGBoost is often the top-performing model for complex data, in this case, its accuracy is lower than both the linear regression and random forest models. This suggests that either the relationships in the data are mostly linear, or that the data does not contain enough complexity or volume for XGBoost to deliver its full potential.

Business Implications:
For this dataset, simpler models performed just as well, or better, than more advanced techniques. The practical takeaway is that for customer sales prediction, a well-tuned linear model or random forest is likely sufficient. If the business brings in new features or additional data, XGBoost may become more useful.

4.6: Model Evaluation: XGBoost Actual vs. Predicted Sales

The XGBoost regression model was evaluated on the test set of 78 customers. Actual average total sales per customer ranged from $832 to $3,013, while the model's predictions ranged from $844 to $3,262. The mean absolute error was $269, and the R^2 score was 0.43, showing that the model explains a moderate amount of the variance in customer sales.

The scatter plot above compares actual and predicted values for each customer. The green dashed line indicates perfect prediction. Most customers are clustered near the prediction line, but a handful show larger discrepancies. The chart includes annotated arrows for the most significant outliers. Goldstar Associates (over-predicted by $756) and Synergy Transfer (under-predicted by $1,181). These outliers highlight where the model struggled, either due to unique customer behavior or factors not captured in the available data.

A closer look at the top ten largest over- and under-predictions shows that errors are distributed across a range of customer types and sales values:

Top 10 Largest Positive Prediction Errors (Over-predicted):

- Customers such as Goldstar Associates, Unity Global, and Sunset Truck Lines had actual sales significantly below what the model forecasted. This suggests they may have recently reduced their business activity, or their historical behavior did not match recent results.

Top 10 Largest Negative Prediction Errors (Under-predicted):

- Customers like Synergy Transfer and Goldstar Global had much higher sales than predicted. These cases may reflect recent growth, unusual sales spikes, or changes in purchasing patterns not previously seen in the training data.

Business Implications:

- The model provides useful, directional insight for most customers but should not be relied on for precise forecasts at the individual account level. Outliers should be reviewed manually, as they may signal accounts in transition or opportunities for targeted outreach.
- Persistent over- or under-prediction for certain customers may indicate missing features in the model (e.g., recent contract wins, lost business, or operational changes). Continuous model retraining and the addition of new data sources could help improve accuracy over time.

Overall, the XGBoost model offers actionable segmentation of the customer base and identifies both high-potential and at-risk accounts for further review.

XGBoost Model: Actual vs. Predicted Sales Plot

- Test set size: 78 customers
- Actual sales range: $832 to $3,013
- Predicted sales range: $844 to $3,262
- Mean Absolute Error: $269.12
- R^2 Score: 0.427

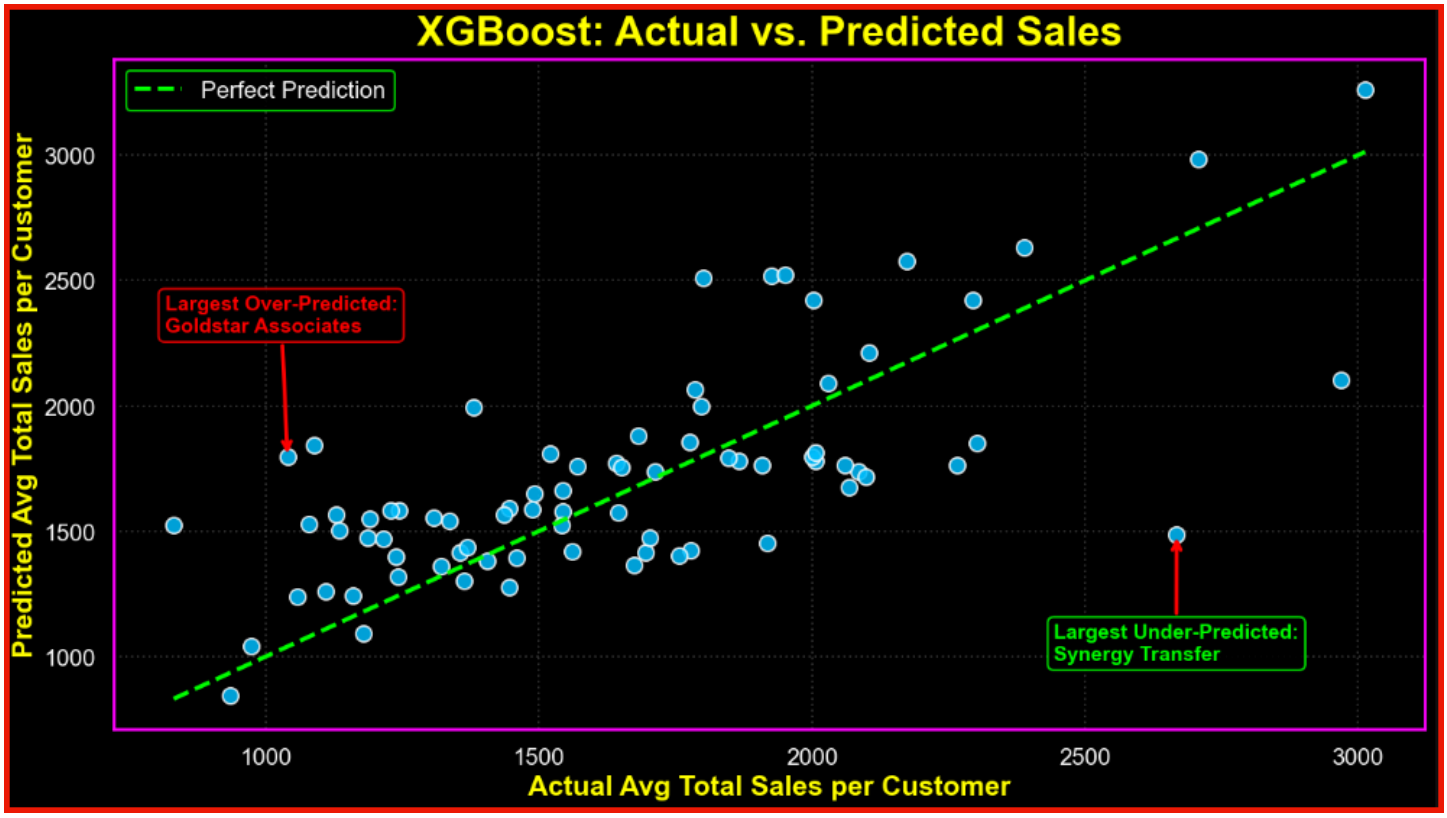
Top 10 Largest Positive Prediction Errors (Over-predicted):

Goldstar Associates | Predicted: $1,796 | Actual: $1,040 | Error: +$756
Unity Global | Predicted: $1,842 | Actual: $1,089 | Error: +$753
Sunset Truck Lines | Predicted: $2,509 | Actual: $1,801 | Error: +$707
Evergreen Corp | Predicted: $1,523 | Actual: $832 | Error: +$691
Evergreen Global | Predicted: $1,995 | Actual: $1,381 | Error: +$614
Stonegate Industries | Predicted: $2,520 | Actual: $1,927 | Error: +$593
Paramount LLC | Predicted: $2,522 | Actual: $1,950 | Error: +$572
Aurora Partners | Predicted: $1,528 | Actual: $1,080 | Error: +$448
Stonegate Tucking | Predicted: $1,564 | Actual: $1,128 | Error: +$436
Summit Solutions | Predicted: $2,421 | Actual: $2,002 | Error: +$419

Top 10 Largest Negative Prediction Errors (Under-predicted):

Synergy Transfer | Predicted: $1,487 | Actual: $2,668 | Error: $-1,181
Goldstar Global | Predicted: $2,104 | Actual: $2,968 | Error: $-864
Sterling Transfer | Predicted: $1,765 | Actual: $2,266 | Error: $-501
Keystone Logistics | Predicted: $1,452 | Actual: $1,918 | Error: $-466

Synergy Industries | Predicted: $1,850 | Actual: $2,303 | Error: $-453
Sunset Tucking | Predicted: $1,674 | Actual: $2,069 | Error: $-395
Aurora Group | Predicted: $1,717 | Actual: $2,099 | Error: $-383
Sterling Group | Predicted: $1,422 | Actual: $1,778 | Error: $-355
Aurora Transfer | Predicted: $1,402 | Actual: $1,756 | Error: $-354
Sterling Associates | Predicted: $1,737 | Actual: $2,085 | Error: $-349



4.7: Model Residuals Analysis: XGBoost Residuals Plot

The chart below shows the residuals for the XGBoost regression model, which predicts each customer's average total sales. Each point represents a customer from the test set. The vertical axis displays the difference between the actual and predicted sales (the residual). A value near zero means the model made an accurate prediction, while larger values (positive or negative) indicate a bigger error.

Most customers are clustered close to the zero line, which means the model is making solid predictions for the majority of accounts. However, a few customers stand out as outliers. The model **under-predicted Synergy Transfer by more than $1,100** (actual sales much higher than predicted), and **over-predicted Goldstar Associates by $750** (predicted sales much higher than actual). These accounts are highlighted directly on the chart.

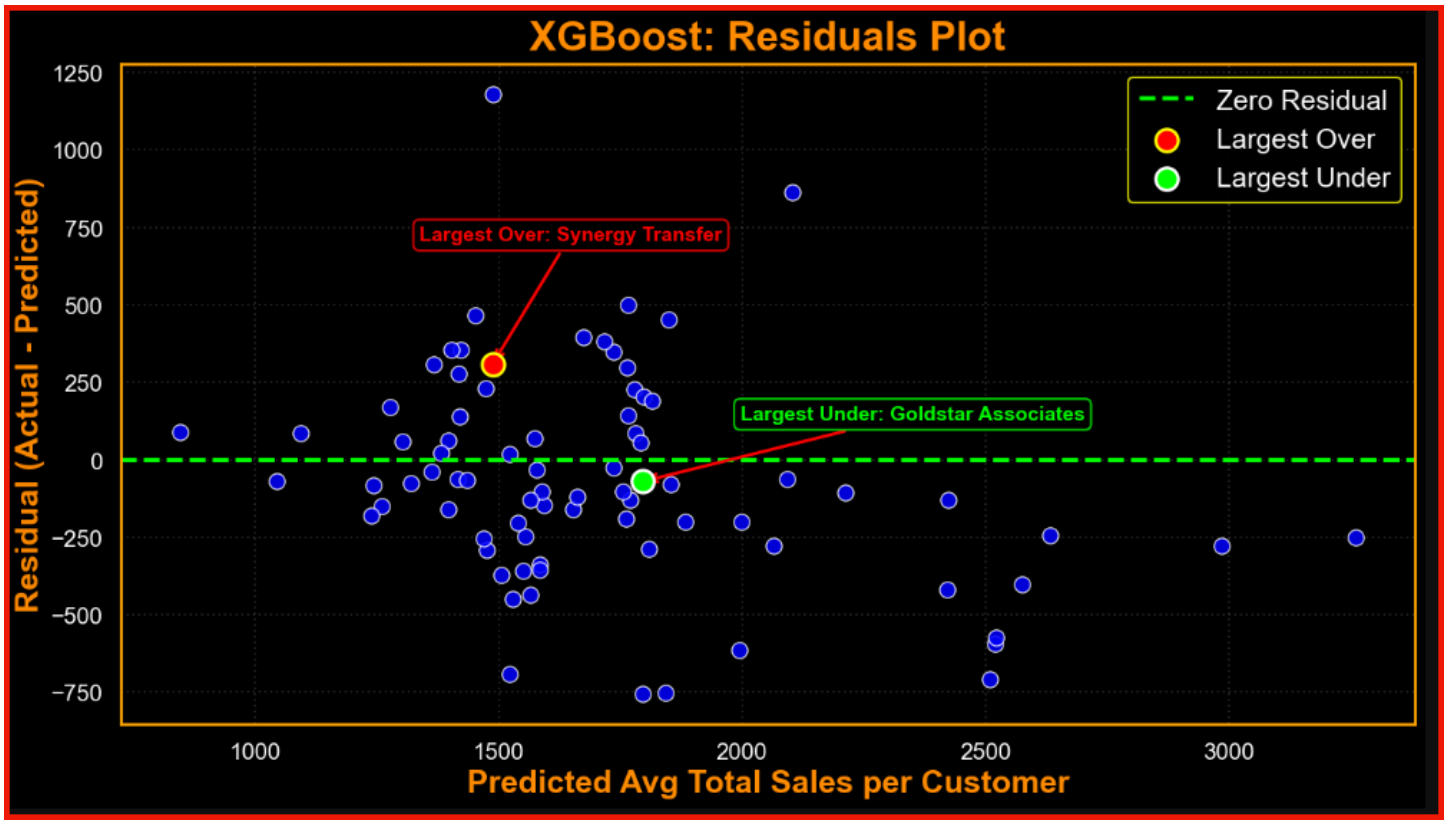
The summary table below the plot lists the ten largest positive and negative residuals. Reviewing these customers can reveal unique business situations, recent changes in purchase patterns, or even data entry issues. Often, the largest errors occur on accounts that behaved very differently during the test period, took on a special project, or made an unusually large or small purchase.

Overall, this residuals analysis confirms that the model works well for most customers. Only a handful of outliers require follow-up or further investigation. The business can use these results to prioritize which accounts to review, and also to guide improvements for future versions of the model.

Key Results:

- **Mean Absolute Error** for the test set is $269.12
- **R^2 score** is 0.427, which reflects moderate predictive power on unseen customers.
- The largest under-prediction is for Synergy Transfer (+$1,181).
- The largest over-prediction is for Goldstar Associates (–$756).
- Most model errors are much smaller and well distributed across the customer base.
- Test set size: 78 customers

Most customer sales predictions are accurate, but the model flags a short list of customers where recent activity did not match historical trends, these outliers offer an opportunity for deeper business review.



4.8: XGBoost Feature Importance Analysis

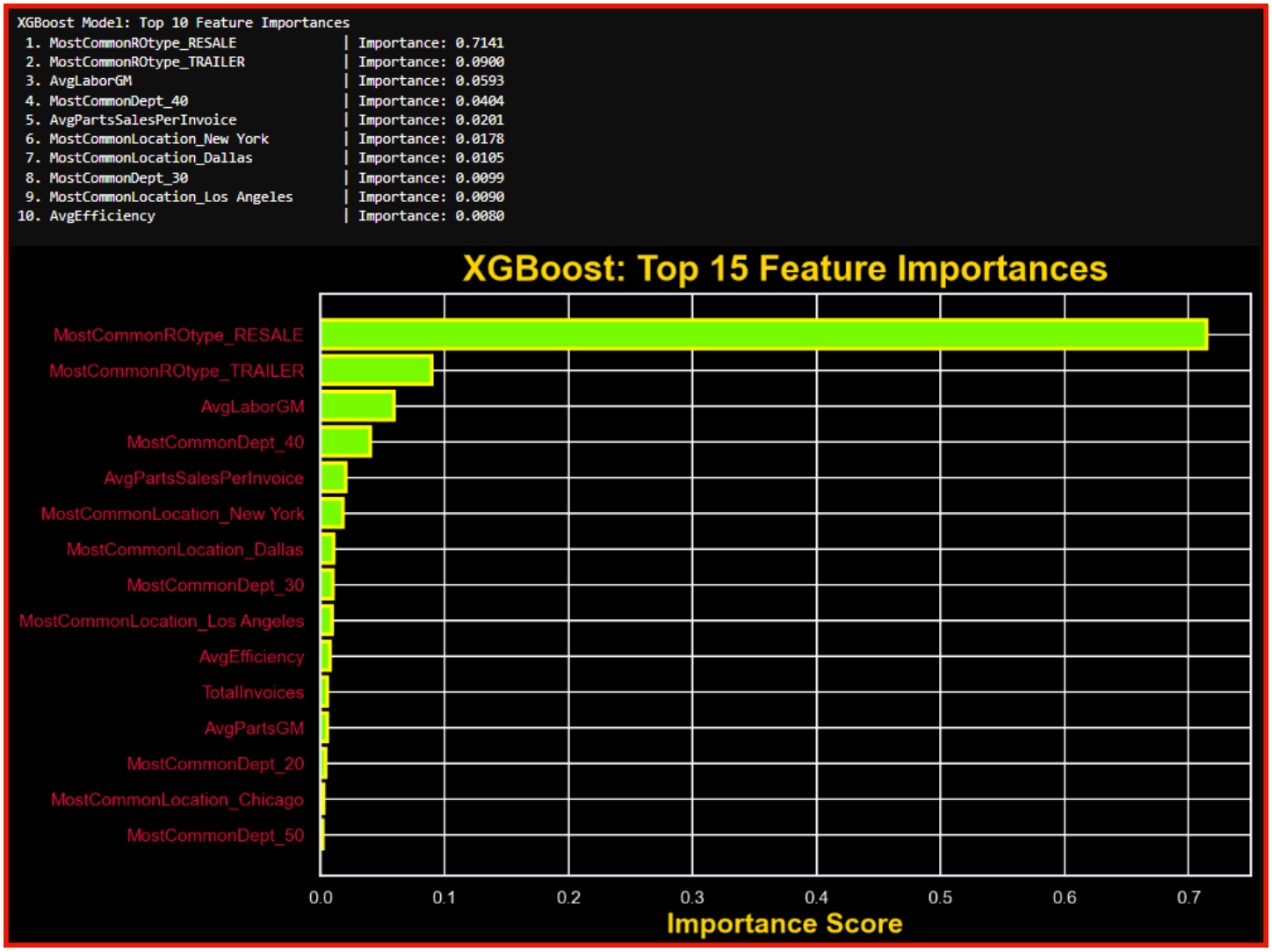
To understand what drives customer value in the model, I extracted feature importances from the final XGBoost regression. The chart and table above show the top features influencing average total sales per customer.

- **The most important factor by a wide margin is whether a customer's most common job type is RESALE.** This aligns with the EDA findings and business intuition: full-service customers (RESALE) generate significantly higher revenue than those focused on counter/parts sales or other job types.
- Other notable drivers include the presence of TRAILER jobs, the customer's average labor gross margin, and their most common department or location.

- Numeric features like average labor GM% and average parts sales per invoice also show up among the top influences, though with much smaller effect compared to categorical business segments.
- Most other variables, including invoice count, efficiency, and parts gross margin, were far less important in the model.

The feature importance results reinforce the idea that customer segmentation, especially by primary service type, matters more for predicting sales than raw volume or efficiency metrics. Dealers looking to grow high-value accounts should focus on increasing RESALE-type work and targeting key departments and locations associated with higher revenue per customer.

This analysis also confirms that the model is finding sensible business patterns and not relying on noise or false predictors.



4.9: Results and Analysis

The modeling phase compared three approaches: linear regression, random forest, and XGBoost. All models were evaluated on a holdout test set of 78 customers, with performance measured by R^2, MAE, and RMSE.

Model performance summary:

- Linear Regression: $R^2 = 0.567$, MAE = $236, RMSE = $302
- Random Forest: $R^2 = 0.525$, MAE = $251, RMSE = $316

- XGBoost: R^2 = 0.427, MAE = $269, RMSE = $348

Across all methods, predictive accuracy was moderate, with linear regression slightly outperforming more complex models. This suggests most relationships in the data are linear or that available features do not strongly support higher model complexity.

Model insights:

- Most predictions were close to actual sales. Residual plots and error tables show most customers cluster near the zero error line, with a few notable outliers.
- Top errors identify accounts for review. The largest under-prediction was Synergy Transfer (actual sales $1,181 above predicted); the largest over-prediction was Goldstar Associates ($756 above predicted). These cases may indicate unique business circumstances or missing features in the dataset.
- Feature importance analysis confirms that customer segment (most common ROtype) is the most powerful predictor of sales. Other features, such as labor gross margin and department, also contribute but with less impact.

Business implications:

- Customer segmentation by job type is critical for understanding value. Dealers should focus on growing and retaining full-service (RESALE) customers.
- Outliers flagged by the model should be manually reviewed. These could be opportunities for additional sales, retention, or risk management.
- Future improvements should focus on collecting more features and more frequent retraining. Adding more granular data on customer behavior, marketing, or external factors may improve accuracy.

Overall, the models provide actionable insights for customer segmentation and account management. The results support data-driven planning for retention, growth, and risk mitigation in dealership operations.

Section 5: Conclusion

This project set out to answer a practical question for dealership management: what distinguishes a high-value customer, and how accurately can we predict average total sales for each account using invoice-level data and basic customer attributes?

By generating a realistic, anonymized invoice dataset, we were able to simulate common business patterns without exposing confidential information. The data included over 45,000 invoices and 387 unique customers, spanning a mix of job types, departments, and locations.

Key findings from the analysis:

- **Customer segmentation is the most powerful driver of value.** The model found that customers whose primary business is RESALE (full-service) work are consistently more valuable than those focused on COUNTER (parts-only) transactions. This aligns with business intuition but was confirmed quantitatively by both EDA and modeling.
- **Invoice volume alone does not guarantee high sales per customer.** Many of the top customers by invoice count were close to the median in sales per invoice, while some of the highest average sales per invoice came from small, one-off accounts.
- **Gross margin and efficiency are less predictive than expected.** While labor gross margin and efficiency are important for profitability, they played a smaller role in predicting customer sales totals

than anticipated. Most customers clustered within a narrow efficiency range, and high gross margins did not always equate to higher total sales.
- **Model accuracy was moderate.** All regression models explained about 43-57 percent of the variance in average customer sales. Linear regression performed slightly better than more complex methods, which suggests that most predictive relationships in this business context are linear or that more granular features are needed for greater accuracy.
- **Model outliers identify real business opportunities and risks.** The largest over- and under-predictions in the test set highlighted accounts whose recent activity diverged sharply from their historical patterns. These cases should be reviewed by management for possible sales opportunities, risk signals, or data issues.

Unexpected findings:

- The most important predictor was not invoice count, efficiency, or even gross margin, but customer segment, as captured by the MostCommonROtype feature. This supports focusing business development on driving more RESALE-type work.
- Despite the large number of invoices and rich feature engineering, adding model complexity did not result in a substantial gain in accuracy. The data did not appear to support strong nonlinear relationships, at least for the features available.
- Some well-known customers (by volume) were not among the most profitable per invoice, underscoring the need for nuanced account management, not just raw sales effort.

Did I get the results I was looking for?

- The project succeeded in quantifying the main business drivers of customer value and provided a practical, reproducible method for segmenting and evaluating the customer base. While the predictive models were not perfect, they offer actionable insight for sales, retention, and operational improvement. The approach is robust and will benefit from further iterations with additional features, real-world data, and periodic retraining.

Next steps:

- Review outlier accounts to determine if business changes, market shifts, or data issues explain the largest errors.
- Collect additional customer features, such as purchase history, customer tenure, or external market data, to improve model performance.
- Use the insights from feature importance to shape sales and marketing strategy, focusing on high-potential customer segments and tailoring offers accordingly.

In summary, this analysis demonstrated that customer segmentation and service mix are the primary levers for driving sales performance in a dealership environment. The tools and methods used here can be readily adapted for ongoing business intelligence and continuous improvement.