FROM ELT PIPELINE TO REGRESSION MODELS: EVALUATING MEDIA MIX MODELS IN PREDICTING NEW CUSTOMERS

Joanne Senoren

D610 Data Engineering Capstone

INTRODUCTION

Name

- Joanne Senoren

Relevant Background

- Full-Stack Web Development
  - Helped debug CRM platforms
  - Works on automation and CRM personalization for email marketing
- Media Strategy and Analytics
  - Managed media advertising budgets for automotive and technology brands
  - Developed digital media plans across search, display, video, and social media
  - Reported on campaign performance using SQL and Tableau
- Thinks marketing is (too often) a shot in the dark.

DIGITAL MEDIA MARKETING CONTEXT

Why This Matters

- Marketers commonly rely on naïve metrics like customer acquisition cost that ignore deeper relationships between spend, seasonality, and performance (Edwards, n.d.).
- Marketing teams move too fast, using fragmented and standardized marketing technology solutions that over-generalizes campaigns, sacrificing effectiveness for speed and perceived efficiency (*Why Traditional Marketing Systems Fail – and How AI Marketing Platforms Like Mowie Lead the Future*, n.d.).
- Benchmarking statistical relationships helps teams:
  - Understand true media effectiveness
  - Allocate budgets more efficiently
  - Forecast new customer acquisitions with data-driven backing
- This study offers a structured ELT, statistical analysis, and modeling workflow that improves measurements beyond typical reporting.

CONTEXT

Project Framework

- E-commerce apparel data across multiple organizations that deployed digital media campaigns over four years.
- New customer acquisitions are count of new customers per day.
- Selected two regression models for testing predictions: OLS and Negative Binomial

RESEARCH AND HYPOTHESIS

Research Question

- How can digital media spend, performance metrics, and seasonality be processed through an ELT (extract, load, transform) solution into a model-ready dataset to evaluate the performance of a Multiple Linear Regression and Negative Binomial Regression model in predicting first purchase acquisitions to support optimizing budget allocation?

Null Hypothesis

- There is no significant relationship between channel media spend, calculated metrics, seasonality, and new customers, and neither models can effectively predict first customer acquisitions.

Alternate Hypothesis

- At least one model includes statistically significant predictors ($p < 0.05$) and demonstrates sufficient explanatory power, defined as an Adjusted $R^2 \geq 0.60$ for multiple linear regression or a Pseudo-$R^2 \geq 0.10$ for Negative Binomial Regression, with 70% of predictions falling within ±20% of the actual values.

AGENDA

1. Data Gathering
2. ELT Architecture Set Up
3. ELT Tasks
4. Exploratory Analysis Summary
5. Model Development | Reduction | Evaluation | Comparisons
6. Model Outcomes
7. Actionable Insights
8. Recommendations
9. Dashboard Overview
10. Expected Outcomes

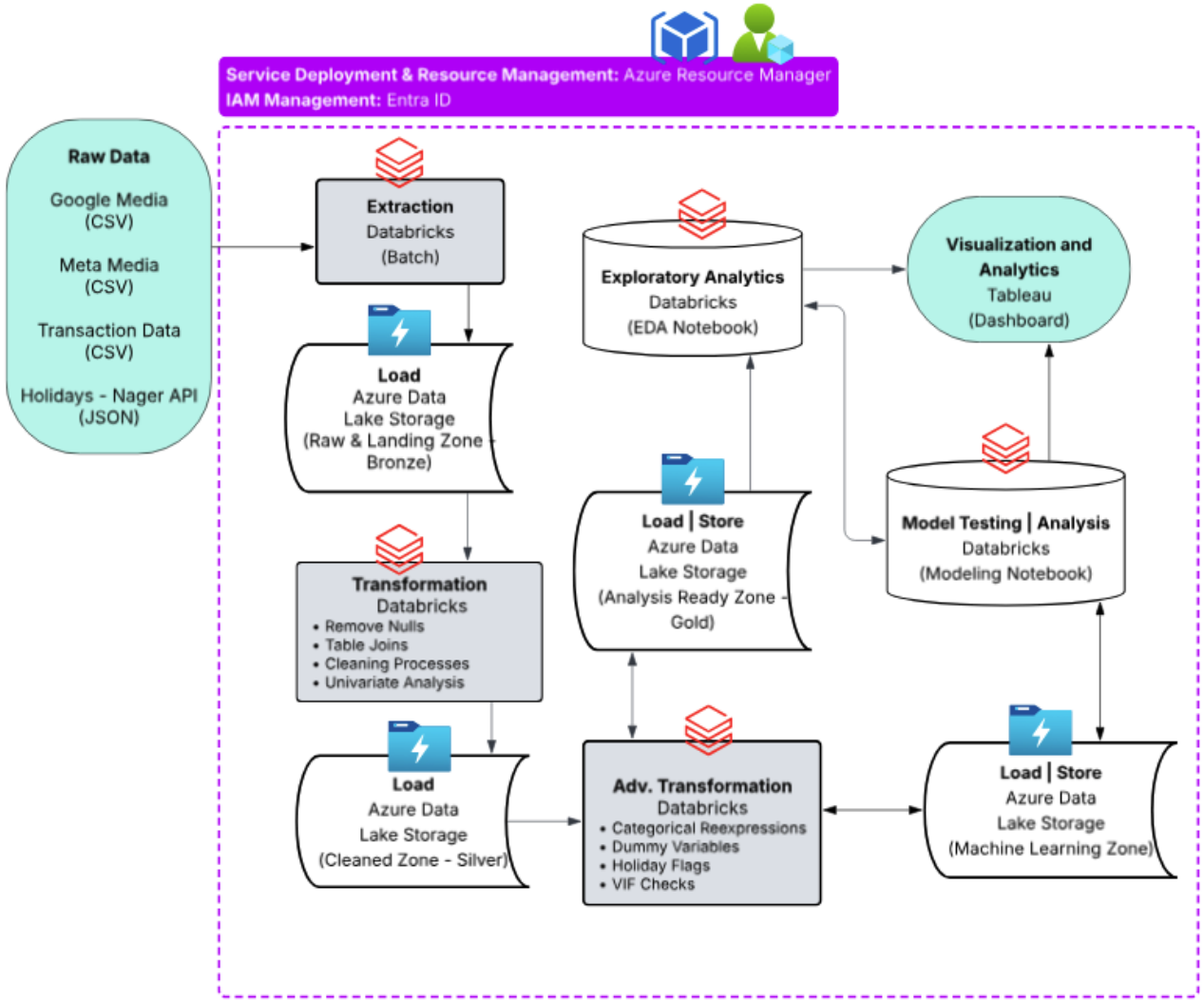
DATA COLLECTION

- Sources:
  - Figshare 2024
    - Real-world digital media placement, anonymized
    - Meta & Google specific placements
      - e.g. google_display, google_shopping
    - Daily data (no spend, spend days), each row is a day
    - Apparel Vertical
  - Nager API
    - For building holiday flags
- Raw Restructuring:
  - Must match vendor reporting format
  - Generate an internal dataset for transactions
  - Composite Key: organization_id + date_day

EXTRACT, LOAD, TRANSFORM APPROACH

Why ELT and not ETL?

- The ELT process is a better solution than ETL because it easily scales to big data (Santos & Costa, 2020).
- Modern marketing data is inherently high volume, high velocity, and high variety, so ELT is the more scalable approach.



MEDALLION HIERARCHY

Overview

- Data design pattern for organizing data in three layers:
  - Bronze (raw)
  - Silver (cleaned)
  - Gold (analysis-ready)
- The ELT process moves data through these medallion stages, saving the progression within the data lakehouse.

```yaml
container: # top-level container
  capstone: # project or workspace name
    bronze: # raw or minimally processed layer
      folders:
        - google # raw Google data
        - meta # raw Meta data
        - internal # raw internal data sources
        - nager_api # raw data from Nager API

    silver: # cleaned, standardized, and conformed layer
      folders:
        - google # cleaned Google data
        - meta # cleaned Meta data
        - internal # cleaned internal data
        - holiday_cleaned # cleaned and expanded holiday data from Nager API

    gold: # analytics-ready or feature-ready datasets
      folders:
        - digital_media # curated digital media datasets for business use
        - ml_prepared # final datasets prepared for machine learning

    machine-learning: # ML-specific outputs and intermediate datasets
      folders:
        - mlr_prepared # data prepared for multiple linear regression models
        - nb_prepared # data prepared for Naive Bayes models
        - nb_pred # prediction outputs from Naive Bayes models
        - ols_pred # prediction outputs from OLS regression models
```

ELT: BRONZE TO SILVER TASKS

(DATA EXTRACT AND CLEANING)

Data Extraction and Staging

- Raw Google, Meta, and internal media CSV files were staged in the Bronze layer.
- Relevant holiday data was fetched from the Nager API, filtered by the media file date range, and stored in Bronze as line-delimited JSON.

Data Quality Cleaning

- Checked for duplicates and null values
- Nulls in media columns across spend, impressions, and clicks (observed on inactive/no-spend days) were imputed as 0 to establish a complete time-series baseline.
  - Row should be zero for spend, impressions, and clicks to be considered inactive

Column Standardization

- Column names were standardized (e.g., all lowercase) to enforce schema conventions
- Columns were renamed for clarity (e.g., 'first purchases' to 'new_customers') based on definitions

ELT: SILVER TO GOLD TASKS

(TRANSFORMATION)

Data Type Transformations

- Converted data types to ensure consistent schema for future analytical processes.
  - Example: Changed date_day from string type to a proper date object.

Holiday Data Enrichment

- Enhanced the holiday dataset by adding fixed holidays not included in the Nager API, but required for complete seasonal analysis.
  - Example: Nager API only provides observed holidays, so if 12/25 is a Sunday, then 12/26 is the API provided date. The fixed holidays like 12/25 was manually added to the holiday list, so the analysis also includes the actual holiday fixed dates.

Feature Augmentation

- Added additional temporal variables such as month, is_weekend, and is_holiday flags to provide richer seasonality signals for analytical models.

Calculated Metric Creation

- Added calculated metrics (e.g., CTR) to enable analytics and reporting teams to quickly access key performance indicators without having to compute them.

Dataset Integration

- Merged multiple datasets into a single dataset and table for streamlined analytics, exploratory analysis, and machine-learning preparation, using a composite key of organization_id and date_day to align records across sources.

ELT: GOLD TO ML TASKS

(ADV. TRANSFORMATION)

Adstock Integration

- Implemented ad decay rates for each channel based on industry benchmarks to account for ad exposure delay over time which is necessary for marketing mix modeling.

Log-Transformations

- Applied log transformations help to normalize skewed data inherent in media performance data and meet multiple linear regression model assumptions.

Dummy Variable Creation

- Converted categorical variables to dummy variables to prepare non-numeric features for linear and negative binomial models.

EXPLORATORY ANALYSIS: WHY?

Objectives during EDA

- Examine both the overall structure and individual characteristics of the dataset.
- Assess the data against key assumptions required for regression modeling.
- Determine the most suitable variables for inclusion in the predictive models.
- Identify trends in new customer acquisition that can support analytics, dashboards, and downstream ELT processes.
  - Compute and review descriptive statistics.
  - Produce univariate and bivariate analyses along with corresponding visualizations.

EDA: DATA HIGHLIGHTS

Correlation With New Customers

- Google PMax showed the strongest correlation with new customer acquisitions, indicating it is the most directly responsive channel in the dataset.
- Other channels demonstrate weaker direct relationships, implying either diminishing returns or less consistent acquisitions.

Channel-Level Acquisition Patterns

- Clear differences appear across channel descriptive statistics, with Google PMax and Google Search delivering the highest average new customers.
  - Google Pmax:

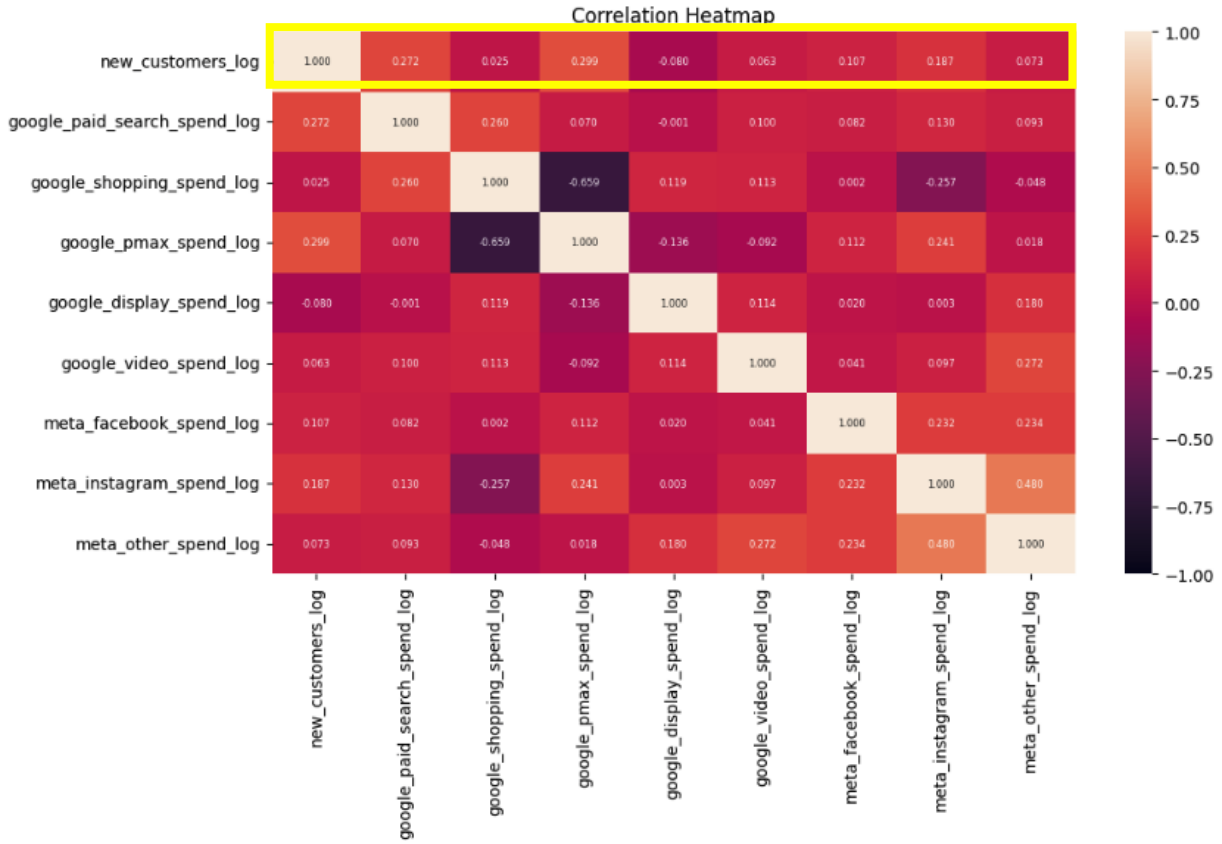
1.2 count	1.2 mean	1.2 std
3824	62.34178870292887	56.648477778057625

  - Google Paid Search:

1.2 count	1.2 mean	1.2 std
5178	57.73271533410583	58.21822241610256

- Meta Facebook showed high average activity (clicks, impressions, CTR), but comparatively weaker correlation with new customers.

	meta_facebook_impressions	meta_facebook_clicks	meta_facebook_ctr_perc
count	8444.000000	8444.00000	8077.000000
mean	43689.881691	893.59711	2.153720



EDA: DATA HIGHLIGHTS

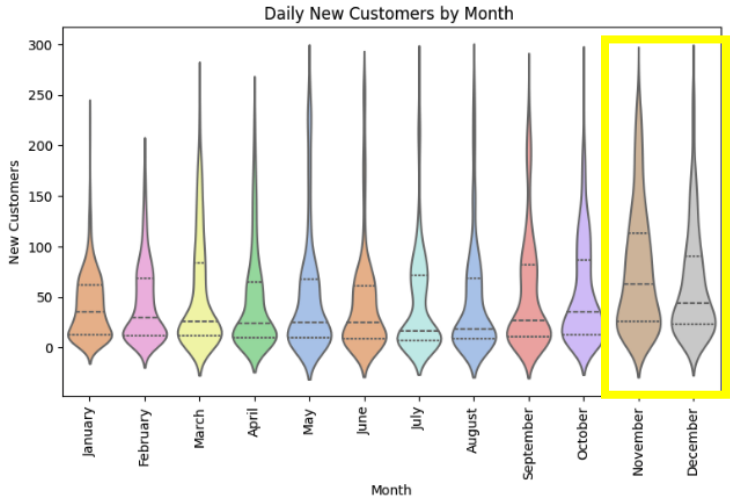
Seasonality Effects

- November and December show the highest average new customer count which implies a seasonal impact.
- Weekend and holiday flags consistently align with at least one new customer, reinforcing temporal influence.
  - Weekend:

weekend_flag	count	mean	std	min
Weekday	5986	51.286835950551286	52.21815702978981	0
Weekend	2387	52.36196062002514	53.69770118293763	1

  - Holiday:

holiday_flag	count	mean	std	min
Holiday	295	54.63389830508475	53.70068410036101	1
Non-Holiday	8078	51.48229759841545	52.604246559024695	0



Active vs. Inactive Media Days

- Active spend days consistently generated 1 new customer, demonstrating the necessity of ongoing media activity.
  - Inactive or no-spend days consistently produced zero new customers.

google_paid_search_no_spend_day	count	mean	std	min
0	5178	57.73271533410583	58.21822241610256	1
1	3195	41.643505477308295	40.12810947359018	0

meta_facebook_no_spend_day	count	mean	std	min
0	8009	52.1975277812461	53.05378624232781	1
1	364	38.29945054945055	40.45125488456249	0

google_display_no_spend_day	count	mean	std	min
0	508	33.38385826771653	34.78868719400089	1
1	7865	52.76948506039415	53.38069749695346	0

meta_instagram_no_spend_day	count	mean	std	min
0	1830	74.49672131147541	71.58013277094189	1
1	6543	45.18752865657955	43.88930397157026	0

google_shopping_no_spend_day	count	mean	std	min
0	5400	49.23777777777778	52.90765174819872	1
1	2973	55.87184661957619	51.89546568375379	0

meta_other_no_spend_day	count	mean	std	min
0	1286	82.65940902021772	76.40217580554678	1
1	7087	45.95611683363906	44.819280469742786	0

google_pmax_no_spend_day	count	mean	std	min
0	3824	62.34178870292887	56.648477778057625	1
1	4549	42.55792481864146	47.17159768024344	0

google_video_no_spend_day	count	mean	std	min
0	355	63.04225352112676	53.30236652713622	6
1	8018	51.086430531304565	52.55956438995091	0

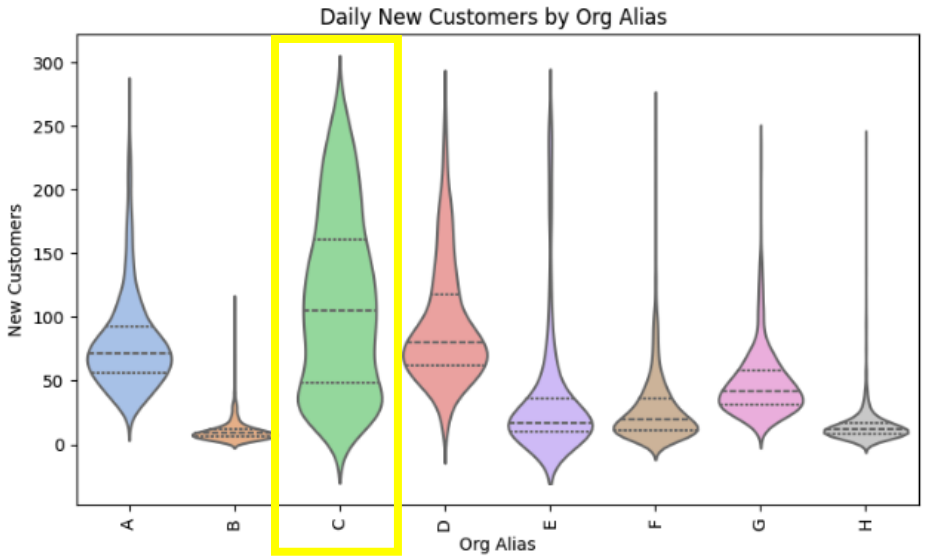
EDA: DATA HIGHLIGHTS

Organizational Performance Patterns

- Organization C produced the highest new-customer counts but with extreme volatility, likely tied to intermittent, high-impact campaigns.
  - Differences in spend behavior and campaign activity across organizations resulted measurable variance in new customers.
  - Organizations grouped by new customers:

	org_alias	count	mean	std	min	25%	50%	75%	max
1	A	1428	79.98109243697479	37.66663504458819	21	56	72	93	270
2	B	1576	10.589467005076141	8.079344038762578	1	6	9	12	113
3	C	998	111.09819639278557	67.83402347625	4	49	105	161	271
4	D	1021	94.06562193927522	47.13710294293667	9	62	80	118	270
5	E	459	36.57298474945534	52.04680260684451	0	10	17	36	264
6	F	1322	27.898638426626324	26.550487770847504	1	11	20	35.75	264
7	G	505	48.853465346534655	27.188026821593244	13	31	42	58	235
8	H	1064	14.8796992481203	14.24759561183901	1	8	12	17	239

- This variance indicates the need to include Organization ID as a model feature to capture inter-organizational effects accurately.



EDA: MODEL PREPARATION TAKEAWAYS

Removing outliers is essential to prevent a few extreme values from overly inflating model coefficients (e.g., in Multiple Linear Regression).

```
count     8444.000000
mean        73.106111
std       1249.310612
min          0.000000
25%         11.000000
50%         33.000000
75%         76.000000
max      96010.000000
Name: new_customers, dt
```

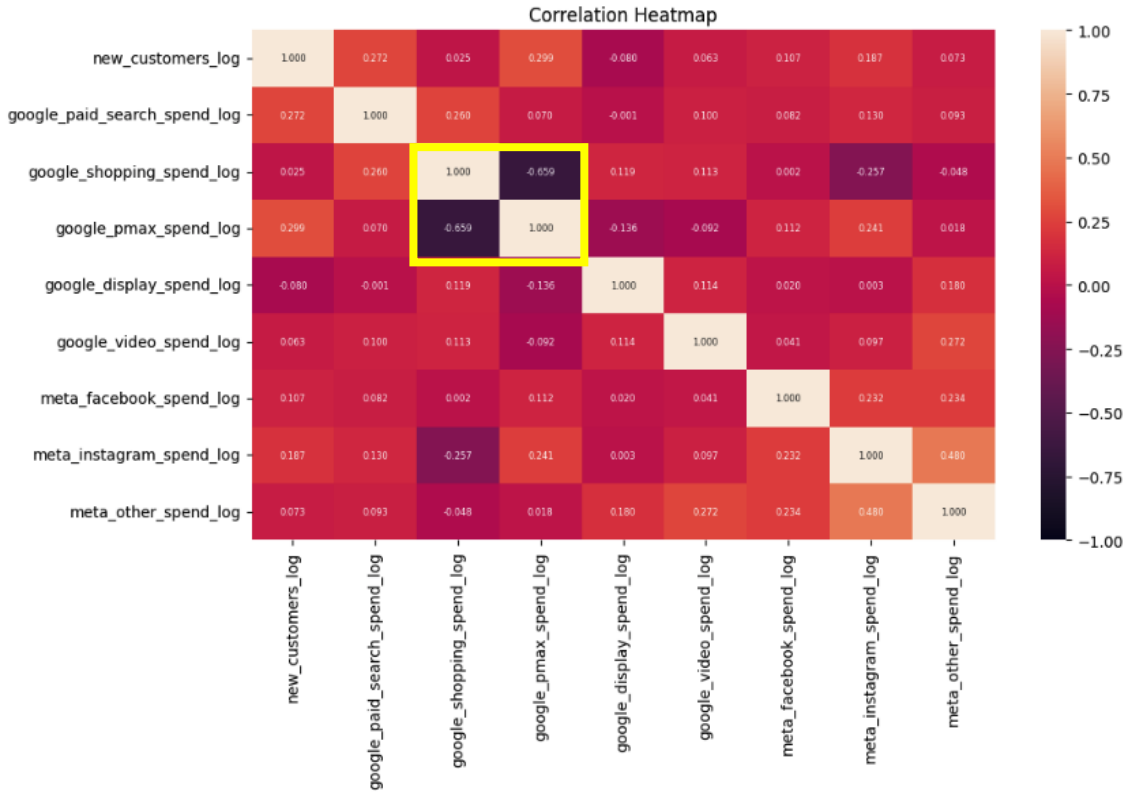
```
# Gets Q1 and Q3 values
q1, q3 = np.percentile(df['new_customers'], [25, 75])

# Calculate interquartile range
iqr = q3 - q1

# Calculate upper limit
upper = q3 + (3 * iqr)
print('Upper Limit:', upper)

# Set variable for count
upper_count = 0

# Count outliers in Population
for x in df['new_customers']:
    if x > upper:
        upper_count += 1

print('Upper Outliers:', upper_count)

range_val = df['new_customers'].max() - upper
print('Upper Outlier Range:', range_val)
```

```
Upper Limit: 271.0
Upper Outliers: 71
Upper Outlier Range: 95739.0
```

```
# Filter out extreme outliers
df = df[df['new_customers'] <= upper]

print(df['new_customers'].max())
```

```
df: pandas.core.frame.DataFrame = [organization_id: object, date
271
```

Collinearity was observed between selected continuous predictor variables in heatmap, necessitating VIF-based reduction to ensure independent and stable model coefficients.

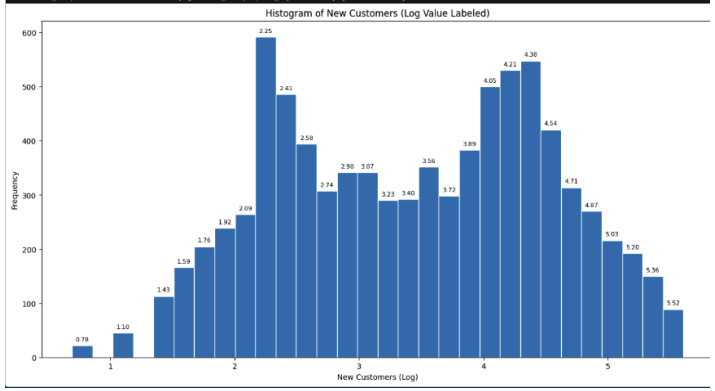


EDA: MODEL PREPARATION TAKEAWAYS

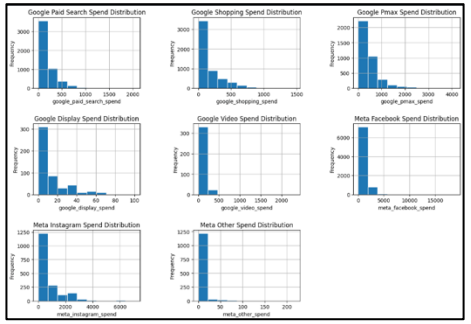
The target variable shows a bi-modal distribution after log-transformation, indicating two distinct campaign groups: highly effective campaigns and generally ineffective campaigns (distinct from zero-spend days).

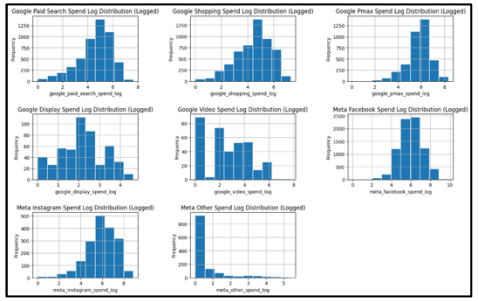
```
# make
media_o

fig, ax

# create histogram
counts, bins, patches = plt.hist(media_on['new_customers_log'], bins=30, edgecolor='white')

# get bin midpoints
bin_mids = 0.5 * (bins[1:] + bins[:-1])

# Add log value labels to get the values of the peaks
for x_mid, y in zip(bin_mids, counts):
    if y > 0:  # only label non-empty bins
        plt.text(x_mid, y + max(counts)*0.01, f"{x_mid:.2f}",
                 ha='center', va='bottom', fontsize=8, rotation=0)

plt.title('Histogram of New Customers (Log Value Labeled)')
plt.xlabel('New Customers (Log)')
plt.ylabel('Frequency')
plt.show()
```

media_on: pandas.core.frame.DataFrame = [organization_id: object, date_day: datetime64[ns] ... 38 more fields]



Log-transformation is required to normalize highly skewed media data and achieve linear relationships between predictors and the target variable (new customers) for the Multiple Linear Regression model.





The 'new customers' target variable exhibits overdispersion, which validates the selection of the Negative Binomial Regression model.

```
# Calculate overdispersion
overdispersion = df['new_customers'].var() / df['new_customers'].mean()
print('Overdispersion:', overdispersion)
```

```
Overdispersion: 53.714437604361486
```

MODEL DEVELOPMENT STEPS

Multicollinearity Check

- Generate Variance Inflator Factors and observe multicollinearity
- Perform VIF-based reduction

Generate test/train sets

- Split the data into test/train sets to check model generalization and predictive performance

Run initial model on train set

- Check Hypothesis Target Metric:
  - Adjusted R2 for Multiple Linear Regression
  - Pseudo-R2 for Negative Binomial Regression

Reduce Model

- Increases statistical significance
- Recursively remove highest p-value above alpha (0.05) variable and re-run model

Generate Predictions

- Use test holdout sets and fit the model
- Compare against actual y-values

Evaluate Errors and Residuals

- Check root mean square error (RMSE) and mean absolute error (MAE)
- Validates model accuracy

Check Model Assumptions

- Examine additional assumptions specific to Multiple Linear Regression and Negative Binomial Regression

SELECTED MODEL FEATURES

Predictor Variables

- Guided by the exploratory analysis and the assumptions of each model, the following variables were selected as suitable predictors.

```
pred_vars = ['google_paid_search_spend_adstock_log','google_shopping_spend_adstock_log', 'google_pmax_spend_adstock_log',
       'google_display_spend_adstock_log', 'google_video_spend_adstock_log','meta_facebook_spend_adstock_log',
       'meta_instagram_spend_adstock_log','meta_other_spend_adstock_log','February','March', 'April', 'May', 'June',
       'July', 'August', 'September','October', 'November', 'December', 'Org_B_9c1f','Org_C_b1a6', 'Org_D_3de0', 'Org_E_1ea2',
       'Org_F_ee4f', 'Org_G_3136','Org_H_4fce', 'is_public_holiday', 'is_weekend']
```

OLS Multiple Linear Regression Target Variable

- **Log** of count of new customers per day

Negative Binomial Regression Target Variable

- **Raw** count of new customer per day

MULTIPLE LINEAR REGRESSION RESULTS

Final Model Summary

- Adjusted R^2 Value is: 0.779
  - The model can explain about 80% of the variance in the data
  - The variables are statistically significant with p-values well below 0.05
- On average, the model's predictions are off by about 18 new customers per day according to mean absolute error (MAE).
- Root mean squared error (RMSE) weighs larger mistakes and is around 31 new customers.
  - The model on some days can have bigger differences between predicted and actual values.
- The model accurately predicted 36.24% of the data within a ±20% margin.
- The QQ Plot residuals plot skewed tails instead of a normal distribution violating the residual normality assumption of multiple linear regression.
  - Predictions on typical days are reasonably reliable, but outliers may under or over-estimate new customers

OLS Regression Results

Dep. Variable:	new_customers_log	R-squared:	0.780
Model:	OLS	Adj. R-squared:	0.779
Method:	Least Squares	F-statistic:	1314.
Date:	Fri, 21 Nov 2025	Prob (F-statistic):	0.00
Time:	21:41:12	Log-Likelihood:	-4949.7
No. Observations:	6698	AIC:	9937.
Df Residuals:	6679	BIC:	1.007e+04
Df Model:	18		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	3.0347	0.032	93.456	0.000	2.971	3.098
google_paid_search_spend_adstock_log	0.0354	0.004	9.854	0.000	0.028	0.042
google_shopping_spend_adstock_log	0.0799	0.004	18.812	0.000	0.072	0.088
google_pmax_spend_adstock_log	0.0634	0.004	17.494	0.000	0.056	0.071
meta_facebook_spend_adstock_log	0.1371	0.005	27.430	0.000	0.127	0.147
meta_instagram_spend_adstock_log	0.0647	0.004	17.492	0.000	0.057	0.072
meta_other_spend_adstock_log	0.0871	0.020	4.435	0.000	0.049	0.126
March	0.1031	0.022	4.649	0.000	0.060	0.147
May	0.1087	0.023	4.761	0.000	0.064	0.154
September	0.1088	0.023	4.647	0.000	0.063	0.155
October	0.1239	0.023	5.388	0.000	0.079	0.169
November	0.4924	0.024	20.276	0.000	0.445	0.540
December	0.2939	0.023	12.666	0.000	0.248	0.339
Org_B_9c1f	-2.1138	0.022	-96.411	0.000	-2.157	-2.071
Org_C_b1a6	-0.5310	0.028	-19.117	0.000	-0.586	-0.477
Org_E_1ea2	-1.7976	0.042	-42.462	0.000	-1.881	-1.715
Org_F_ee4f	-1.1800	0.020	-59.272	0.000	-1.219	-1.141
Org_G_3136	-0.8636	0.034	-25.659	0.000	-0.930	-0.798
Org_H_4fce	-2.1271	0.024	-88.471	0.000	-2.174	-2.080

Omnibus:	176.915	Durbin-Watson:	2.027
Prob(Omnibus):	0.000	Jarque-Bera (JB):	432.837
Skew:	-0.031	Prob(JB):	1.02e-94
Kurtosis:	4.244	Cond. No.	61.2

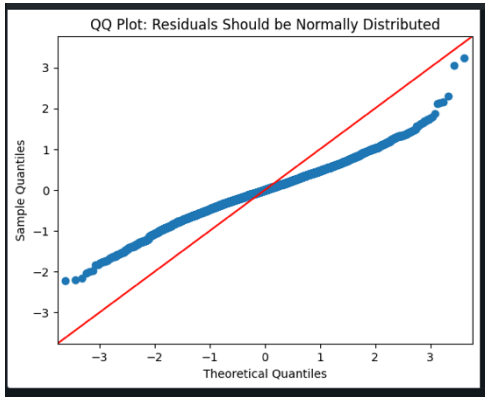
Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

```
# Actual and Predicted
y_true = comparison_df['Actual']
y_pred = comparison_df['Predicted']

# RMSE
mse = mean_squared_error(y_true, y_pred)
rmse = np.sqrt(mse)

# MAE
mae = mean_absolute_error(y_true, y_pred)
mae = np.round(mae)

print(f"RMSE: {rmse:.2f}")
print(f"MAE: {mae:.2f}")
```

```
RMSE: 30.73
MAE: 18.00
```

	Actual	Predicted
5747	15	15
1734	100	127
8103	7	8
6089	21	28
6838	108	130
4889	29	34
4425	2	7
1655	243	132
8235	6	7
7187	119	109

Proportion of predictions within ±20% of actual counts: 36.24%



NEGATIVE BINOMIAL REGRESSION RESULTS

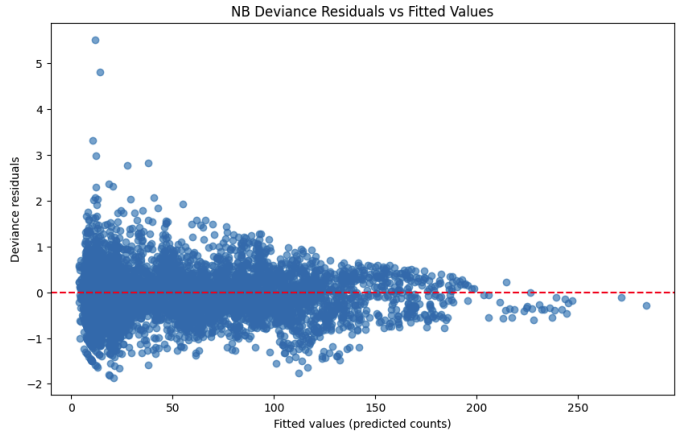
Final Model Summary

- Pseudo-R^2 Value is: 0.5545
  - The model is 55% better than a null model
  - The variables are statistically significant with p-values well below 0.05
- On average, the model's predictions are off by about 19 new customers per day according to mean absolute error (MAE).
- Root mean squared error (RMSE) weighs larger mistakes and is around 31 new customers.
  - The model on some days can have bigger differences between predicted and actual values.
- The model accurately predicted 34.59% of the data within a ±20% margin.
- The deviance residuals scatterplot show a patterned shape with a larger number of errors from the left that tapers off to the right.
  - The heteroscedastic shape indicates a poor model fit.

Generalized Linear Model Regression Results

Dep. Variable:	new_customers	No. Observations:	6698
Model:	GLM	Df Residuals:	6678
Model Family:	NegativeBinomial	Df Model:	19
Link Function:	Log	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-30440.
Date:	Fri, 21 Nov 2025	Deviance:	1841.4
Time:	06:26:35	Pearson chi2:	2.61e+03
No. Iterations:	11	Pseudo R-squ. (CS):	0.5545
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	3.1783	0.066	48.020	0.000	3.049	3.308
google_paid_search_spend_adstock_log	0.0342	0.007	4.747	0.000	0.020	0.048
google_shopping_spend_adstock_log	0.0795	0.009	9.311	0.000	0.063	0.096
google_pmax_spend_adstock_log	0.0541	0.007	7.448	0.000	0.040	0.068
meta_facebook_spend_adstock_log	0.1232	0.010	12.160	0.000	0.103	0.143
meta_instagram_spend_adstock_log	0.0611	0.007	8.182	0.000	0.046	0.076
meta_other_spend_adstock_log	0.1284	0.039	3.255	0.001	0.051	0.206
March	0.1210	0.045	2.671	0.008	0.032	0.210
May	0.1198	0.047	2.565	0.010	0.028	0.211
June	0.1149	0.052	2.215	0.027	0.013	0.217
September	0.1397	0.048	2.927	0.003	0.046	0.233
October	0.1478	0.047	3.154	0.002	0.056	0.240
November	0.5891	0.049	12.026	0.000	0.493	0.685
December	0.4385	0.047	9.346	0.000	0.347	0.530
Org_B_9c1f	-2.1675	0.045	-48.681	0.000	-2.255	-2.080
Org_C_b1a6	-0.4204	0.055	-7.578	0.000	-0.529	-0.312
Org_E_1ea2	-1.8095	0.086	-21.051	0.000	-1.978	-1.641
Org_F_ee4f	-1.1511	0.040	-28.853	0.000	-1.229	-1.073
Org_G_3136	-0.8186	0.067	-12.167	0.000	-0.950	-0.687
Org_H_4fce	-2.0753	0.049	-42.655	0.000	-2.171	-1.980




```python
# Actual and Predicted
y_true = comparison_df['Actual']
y_pred = comparison_df['Predicted']

# RMSE
mse = mean_squared_error(y_true, y_pred)
rmse = np.sqrt(mse)

# MAE
mae = mean_absolute_error(y_true, y_pred)
mae = np.round(mae)


print(f"RMSE: {rmse:.2f}")
print(f"MAE: {mae:.2f}")
```

```
RMSE: 30.61
MAE: 19.00
```

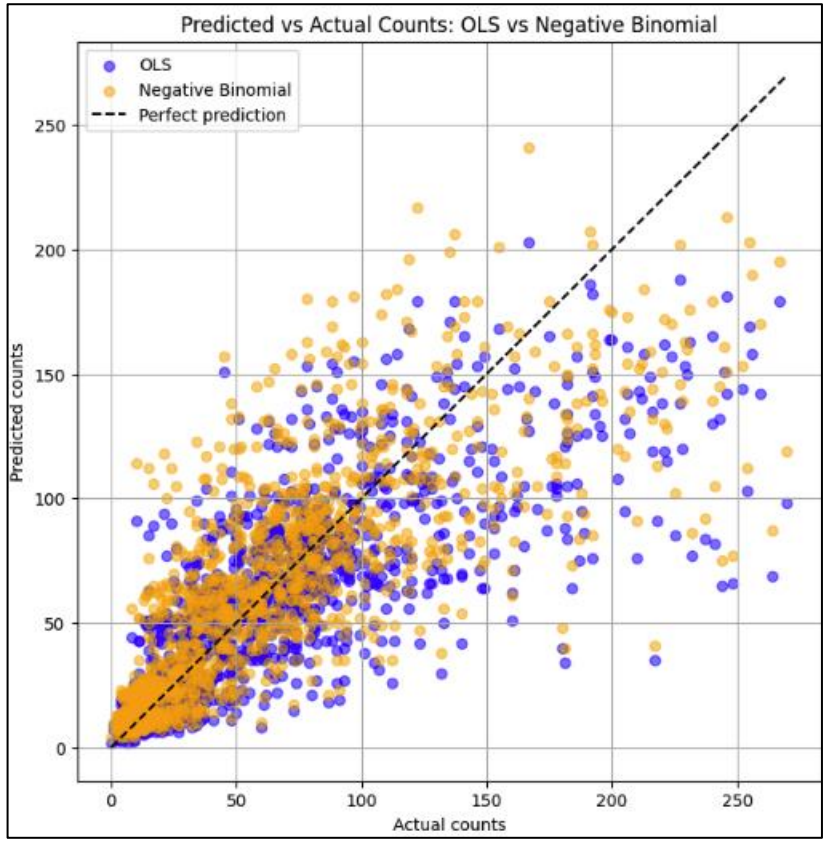
Sample of predictions vs actual:

	Actual	Predicted
5747	15	19
1734	100	154
8103	7	9
6089	21	33
6838	108	136
4889	29	41
4425	2	9
1655	243	145
8235	6	8
7187	119	127

Proportion of predictions within ±20% of actual counts: 34.59%

MODEL COMPARISON

- The OLS Multiple Linear Regression analysis MAE is slightly better than the Negative Binomial Regression
  - 18.00 (OLS) < 19.00 (NB)
    - On average, the OLS MAE is one less than Negative Binomial
- The Negative Binomial Regression RMSE is barely better than the OLS Multiple Linear Regression Model
  - 30.61 (NB) < 30.73 (OLS)
    - Based on RMSE which penalizes extreme errors
- The OLS prediction coverage within ±20% is higher than Negative Binomial
  - 36.24% (OLS) > 34.59% (NB)
    - OLS has a higher predictive ability than Negative Binomial
- Both models tend to struggle with predicting extremely large new customer counts due to few sporadic, high activity campaigns.



MODEL OUTCOMES

Hypotheses Review

- H_0: "There is no significant relationship between channel media spend, calculated metrics, seasonality, and first purchases, and neither models can effectively predict first customer acquisitions."
- H_1: "At least one model includes statistically significant predictors ($p < 0.05$) and demonstrates sufficient explanatory power, defined as an Adjusted $R^2 \geq 0.60$ for Multiple Linear Regression or a Pseudo-$R^2 \geq 0.10$ for Negative Binomial Regression, with 70% of predictions falling within ±20% of actual values."

Findings

- Both MLR and Negative Binomial models identified statistically significant predictors and showed strong explanatory power.
- Neither model met the 70% prediction accuracy threshold (only 34–35% coverage), failing to predict extreme spikes in customer counts.
- Analysis result is a mixed outcome: The null hypothesis must be rejected because the models have highly statistical coefficients, but the alternate hypothesis cannot be fully accepted because neither models met the prediction coverage threshold of 70%.

Insights

- Models handle no-spend days well, capturing baseline inactivity and accounting for better predictions in lower customer counts.
- Models provide high descriptive value, but limited reliability for precise new customer forecasting.

Limitations

- Challenges include skewed distributions, extreme values, and spikes in new customer counts.

ACTIONABLE INSIGHTS

Use exploratory data analysis (EDA) insights and significant predictors for strategy and budgeting.

Leverage OLS model's high explanatory power to support budget recommendations for media planning, highlighting seasonality and channel performance.

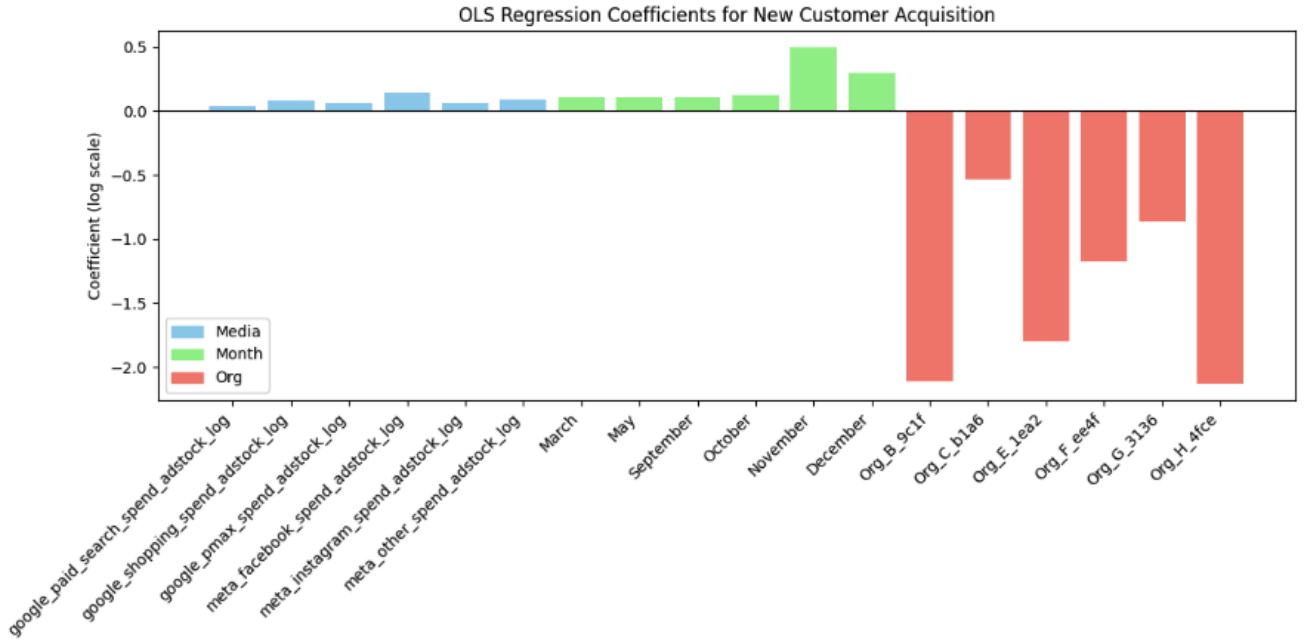
Gather data for enrichment (creative messaging, offline spend, competitive data, weather, additional organizations) which could improve predictive capabilities for future models

Do not use the models for exact new customer forecasting.

MEDIA RECOMMENDATION

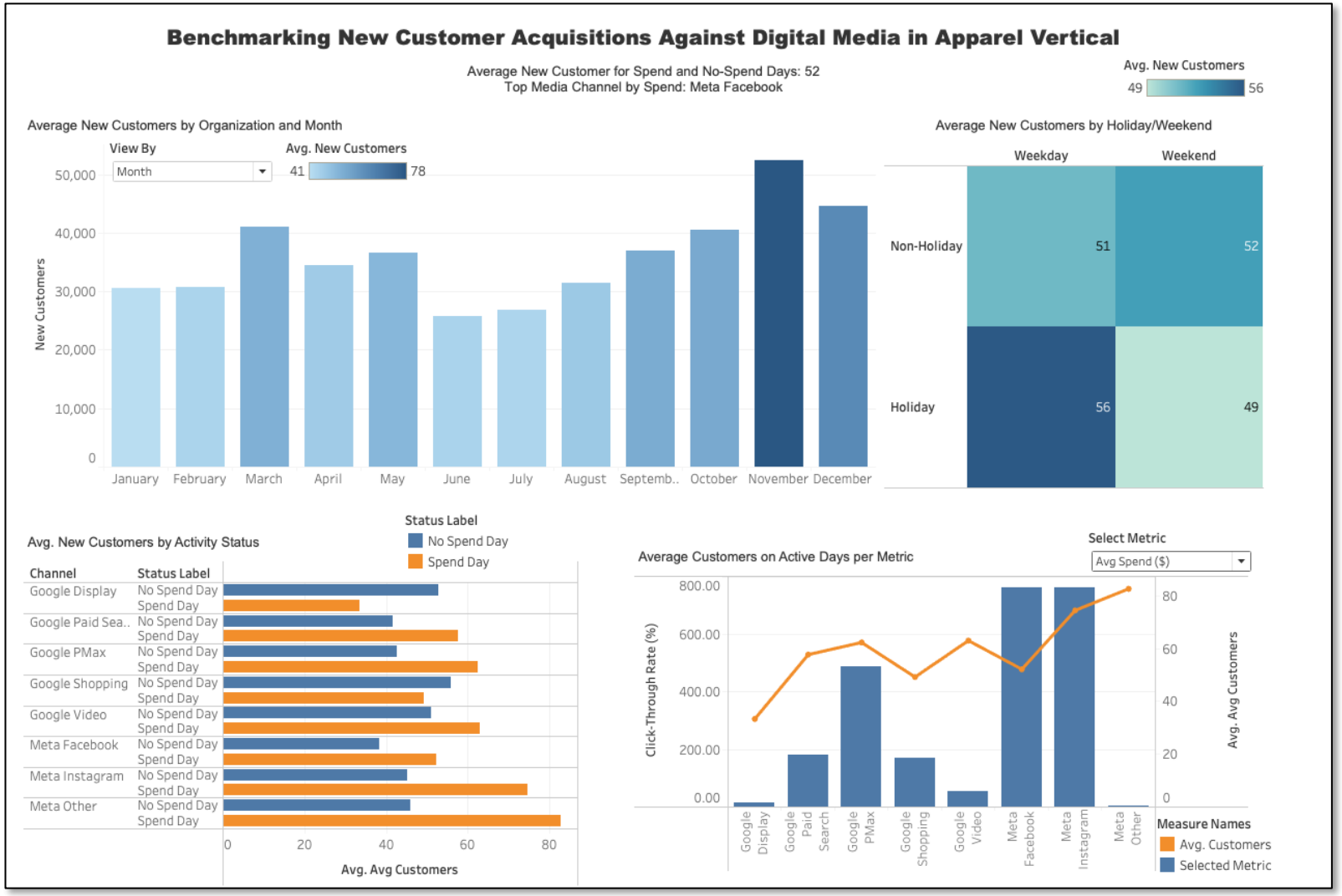
Top Drivers of First Customer Acquisitions Based on OLS Model

- **High-impact channels:** Meta Facebook and Google Shopping show strongest positive effects.
- **Seasonality:** October–December have peak acquisition; align campaigns with these periods.
- **Organizational performance:** Focus high-effort campaigns organizations closer to the baseline (smaller negative coefficients), consider low-cost strategies for weaker performers.



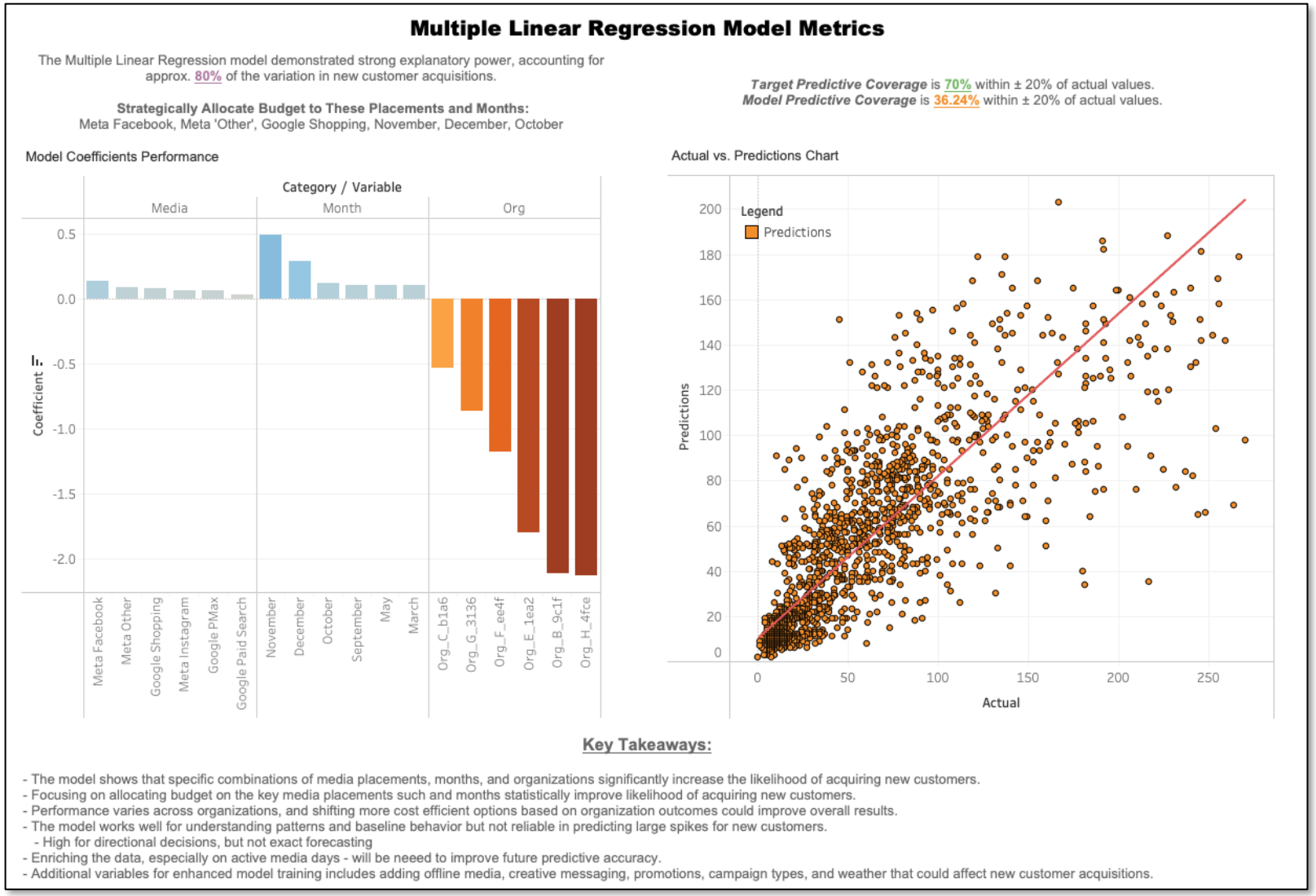
BUSINESS REPORTING DASHBOARD

- Visualization component of ELT pipeline.
- [Tableau Public Link]



BUSINESS REPORTING DASHBOARD

- Visualization component of ELT pipeline.
- Tableau Public Link



EXPECTED BENEFITS

Improving Strategic Planning and Media Budget Allocation

- The selected OLS Multiple Linear Regression model explains about 80% of variation in new customers through its results.
- While not ready for precise volume forecasting, the model reliably identifies directional lift associated with spend and seasonality.
  - e.g. A 1% increase in Meta Facebook spend is associated with an approximate 0.137% increase in new customer acquisitions, holding all other variables constant.
- Supports more confident planning for high-value periods (e.g., November + December producing the highest customer averages).
- Teams can assertively reallocate media spend from lower-performing channels/organizations to higher-performing ones, improving target metric of new customers without increasing total budget.
- Model provides quantitative evidence of which variables truly impact new customer acquisitions, rather than relying on intuition or naive metrics.

Establishing a Foundation for More Advanced Predictive Modeling

- Data enrichment recommendations focusing on adding 5 target variables such as creative messaging, offline spend, promotions, weather, and competitive data create a roadmap for ongoing accuracy improvements.

Scaled and Streamlined Workflow by Leveraging ELT Solution in Azure

- Automating ELT process in Azure environment have proven to reduce manual data processing by 60% ("Azure-Powered Marketing Data Centralization," 2025) .
- Saving data in Delta format on Azure Data Lake Storage improves update/merge operation performance by up to 56% (Harris, n.d.).
- Leveraging serverless DLT on top of Spark in Databricks saves compute costs of up to 98% (Lappas et al., 2024).

Enhanced Reporting and Visibility with Dashboards

- Gold-layer data and dashboards provide consistent and automated reporting.
- Reduces manual data visualization preparation time for marketing teams, freeing time for strategic planning.

REFERENCES

1. Azure-Powered Marketing Data Centralization. (2025, November 12). *mu-sigma*. Retrieved November 28, 2025, from https://www.mu-sigma.com/case-study/azure-powered-marketing-data-centralization/
2. Edwards, S. (n.d.). *CAC is Lying to You*. https://marketer.co/blog/cac-is-lying-to-you
3. Harris, J. (n.d.). *Delta Lake Performance | Delta Lake*. Delta Lake. Retrieved November 28, 2025, from https://delta.io/blog/delta-lake-performance/
4. Lappas, P., Armbrust, M., Dalwadi, M., Neumann, A., Liang, X., Kianfar, K., & Wang, K. (2024, August 27). *Cost-effective, incremental ETL with serverless compute for Delta Live Tables pipelines | Databricks Blog*. Databricks. Retrieved November 28, 2025, from https://www.databricks.com/blog/cost-effective-incremental-etl-serverless-compute-delta-live-tables-pipelines
5. *Why traditional marketing systems fail – and how AI marketing platforms like Mowie lead the future*. (n.d.). https://www.mowie.ai/blogs/why-traditional-marketing-systems-break-down-and-how-ai-bridges-the-gap