STRATUM REGIONAL MODELS

8/1/2022

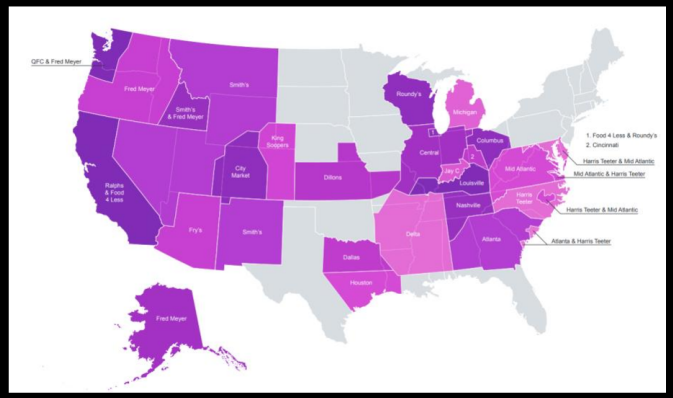
Prathamesh Kulkarni

Decision and Data Science Intern

2022 - D&DS INTERN PROJECT STATEMENT

PRISM

- **Outcome:** Develop reproducible and scalable Machine learning capabilities to shape GPCB go-to-market strategies specific to regional pricing/assortment recommendations
- Develop machine learning models using Kroger Division level data to help drive sales. Provide regional recommendation to the business on how to increase GP's share of volume
- **Scope**: Bath Tissue and Paper Towel
- **Deliverables:**
  - Mathematical models and visualizations (Alteryx, Tableau, documentation, etc.)
  - Insights and recommendation for pricing, promotion and/or assortment changes
  - Recommendations for possible Market Test and Learn opportunity
- **Stakeholders**: Sales, Category Management, Brand, Customer Investments, Data and Decision Science team
- **Timelines:**
  - Internship Period May 17th – Aug 5th
  - Final Presentation – Aug 1st



BUILDING THE MODELS

Key learnings along the way:

- Need to remove outliers that are affecting the % discount variable.
- Needed to run model (PPG level) having minimum of a 1.5 years data to have moderate model fit and coefficient values.
- The stratum data is highly correlated that needs to undergo heavy transformation (Log, Power, square root) to make.

PROJECT SCOPE :-

PRISM

- Total 450 Models (PPG level)
- Main predicting variables from two different Kroger data sources:-

  1) Market 6: Discount %, AD %, Display %, Price

  2) Stratum: Income Group (High, Med, Low) , Age Group (High, Med, Low), Gender – (Male, Female)

  3) Base price is a calculated using a separate algorithm.
- Volume predicted and interacting variable analyzed.
- Income group are bifurcated as:-

  1) High Income group- 100 – 150 K

  2) Medium Income group- 50- 100 K

  3) Low Income group- up to 50 K
- Age group has been defined as :-

  1) High Age group:-65 - 75

  2) Medium Age group:- 35- 65

  3) Low Age Group:- 21- 35
- GP products as well as competitor products included in modelling.
- Ridge Regression modelling selected as the best modelling technique.
- Analytics Dashboard made using various charts.
- Regions Included are :- USA, Georgia, Indiana, Ohio, Kansas ,Illinois, California , Kentucky, Michigan, Tennessee, Texas, Virginia, Wisconsin, Washington
- Creating a dummy pipeline using the AWS environment for one regional model.

OUTPUT INTERPRETATION

- $Volume = constant + a * AD\% + b * Display\% + c * Stores\ selling\ count + d * Sales\ by\ Male + e * Sales\ by\ Female + f * Sales\ by\ low\ income + g * Sales\ by\ med\ income + h * Sales\ by\ high\ income + i * Sales\ by\ low\ age + j * Sales\ by\ med\ age + k * Sales\ by\ high\ age + l * Sales\ by\ Base\ Price + m * Sales\ by\ \%\ Discount + n * Sales\ by\ 1st\ Week\ of\ month + o * Sales\ by\ 2nd\ Week\ of\ month + p * Sales\ by\ 3rd\ Week\ of\ month + q * Sales\ by\ 4th\ Week\ of\ month + r * Sales\ by\ 5th\ Week\ of\ month$

Division	State	Stratum PPG encoded	Const	R2_Score	MAE %	AD %	Display %	STORES_St	Vol-Male	Vol-female	income_low	income_mec	income_high	age_low	age_med	age_high	Base Price
Atlanta	Georgia	BRAWNY : 12 EQ : 6 CT : EXTRA LARGE : UNSCENTED	-0.0362	0.66197	0.00241	0.0019	0.02147	0.07338	0	-3.59E-33	-7.17E-33	0	-4.59E-31	0	-1.43E-32	-2.30E-31	-0.0415
Atlanta	Georgia	CHARMIN : 36 EQ : 9 CT : 4X : UNSCENTED	0.12812	0.61482	0.01116	0.00873	0.09764	0.01039	0.00044	0.0147574	-0.009599	0.0081554	-0.034780403	-0.00191	-0.02218	-0.00321	-0.21333

R2 is used to determine the efficiency of the variable in explaining the output.

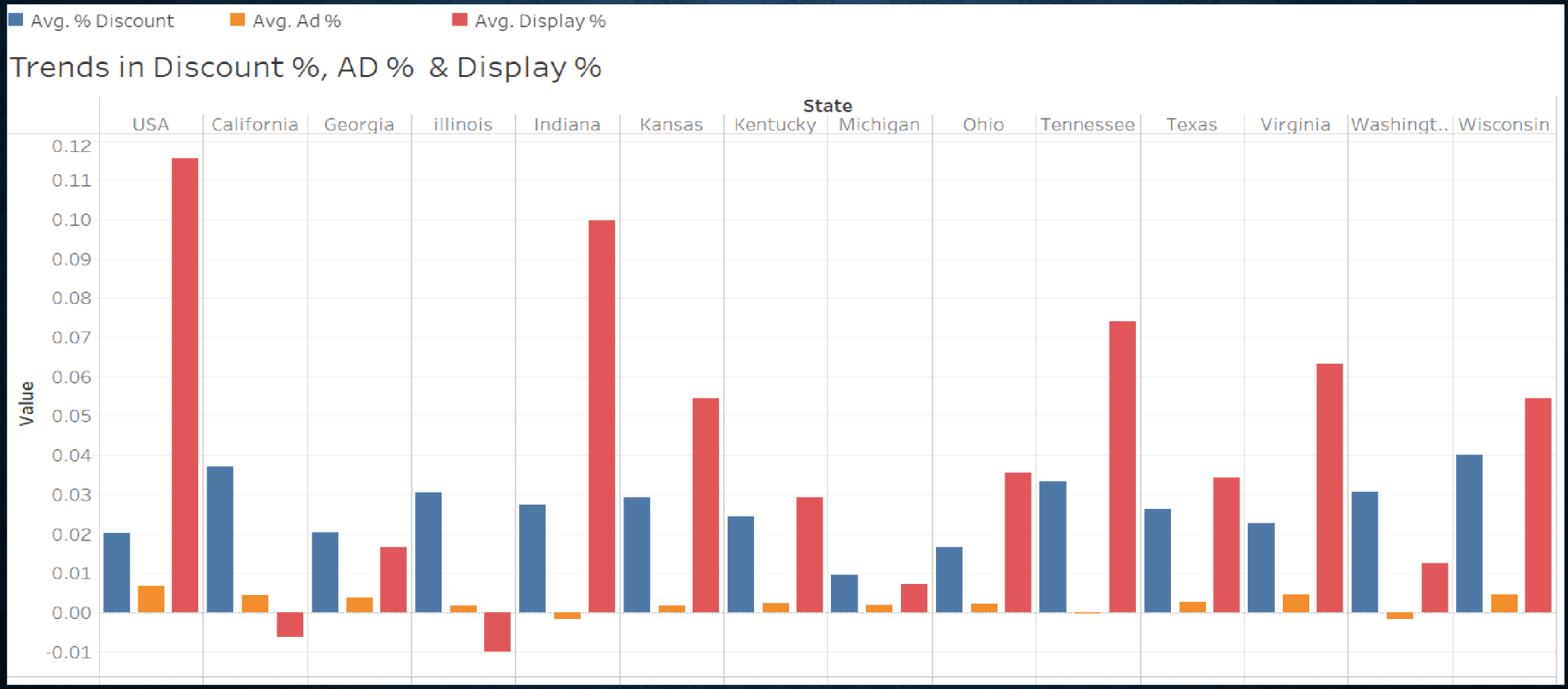
MAE % Is used to represent the percentage error in predicted & actual value of output.

We can say that for 1 unit increase in sales by Male there is o unit addition in sales.

We can say that for 1 unit increase in sales by high income group there is decrease of 0.034 units is sales..

MODEL RESULT

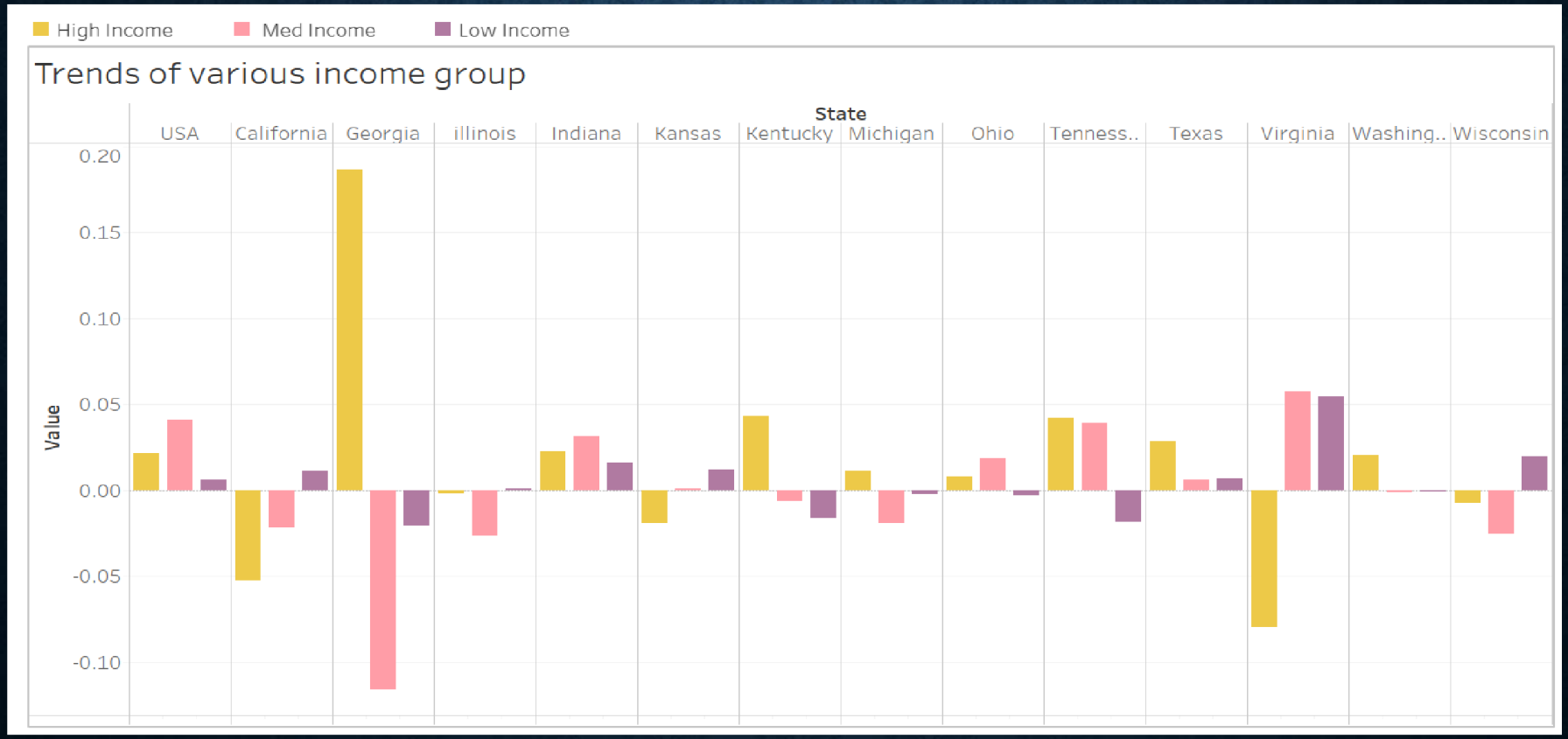
A clear trend that we can see is that display are generally more contributing to Sales rather than Ad and discount given.



MODEL RESULT

A clear trend we can see that medium income group is generally more influential buying the products with exception of state of Georgia. While low-income group performed better in the state of Wisconsin.

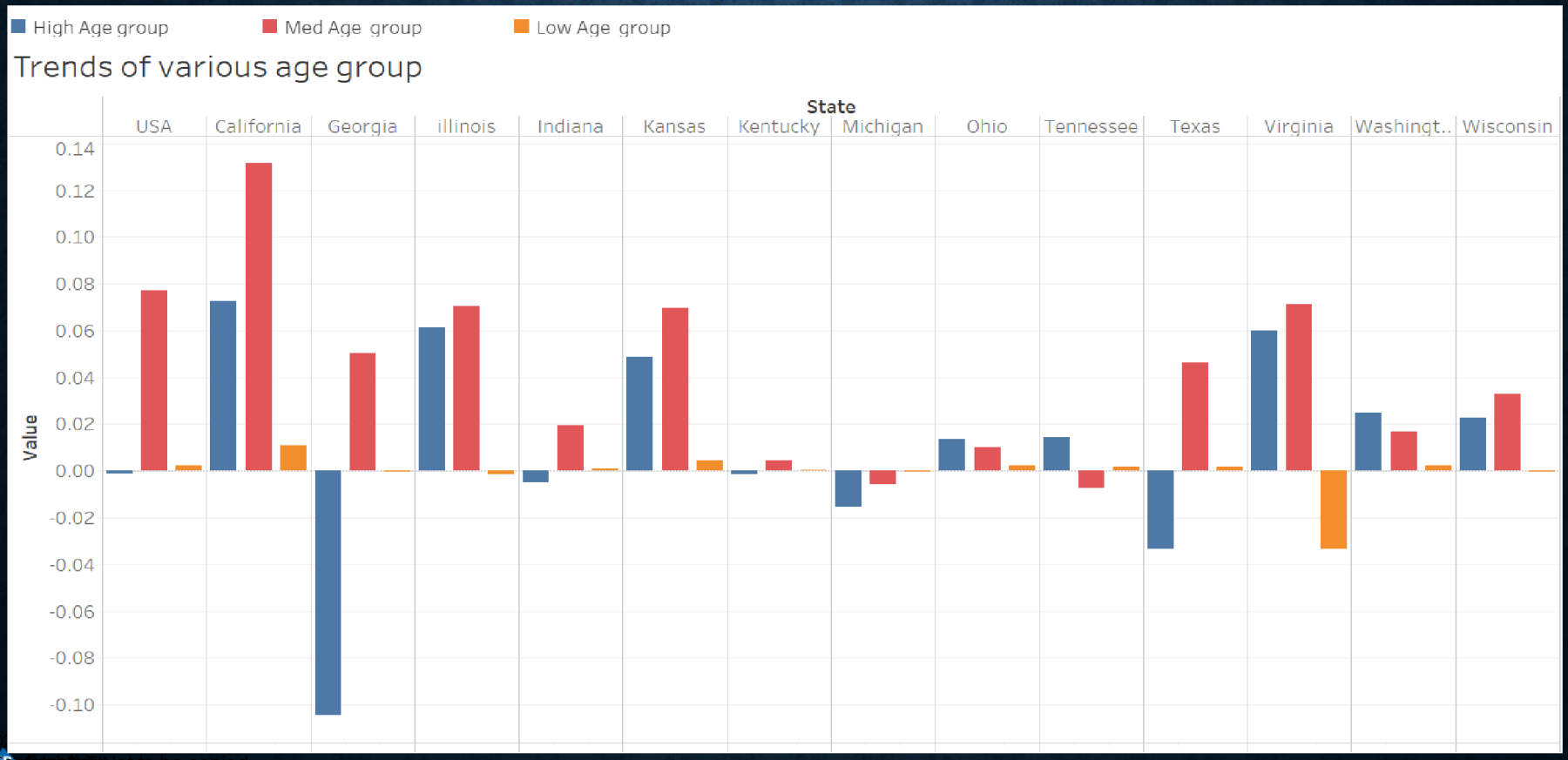
PRISM



MODEL RESULT

A clear trend is that Med age group are more influential in buying the products followed with high age group people, while low age group don't contribute much to the sales of the products.

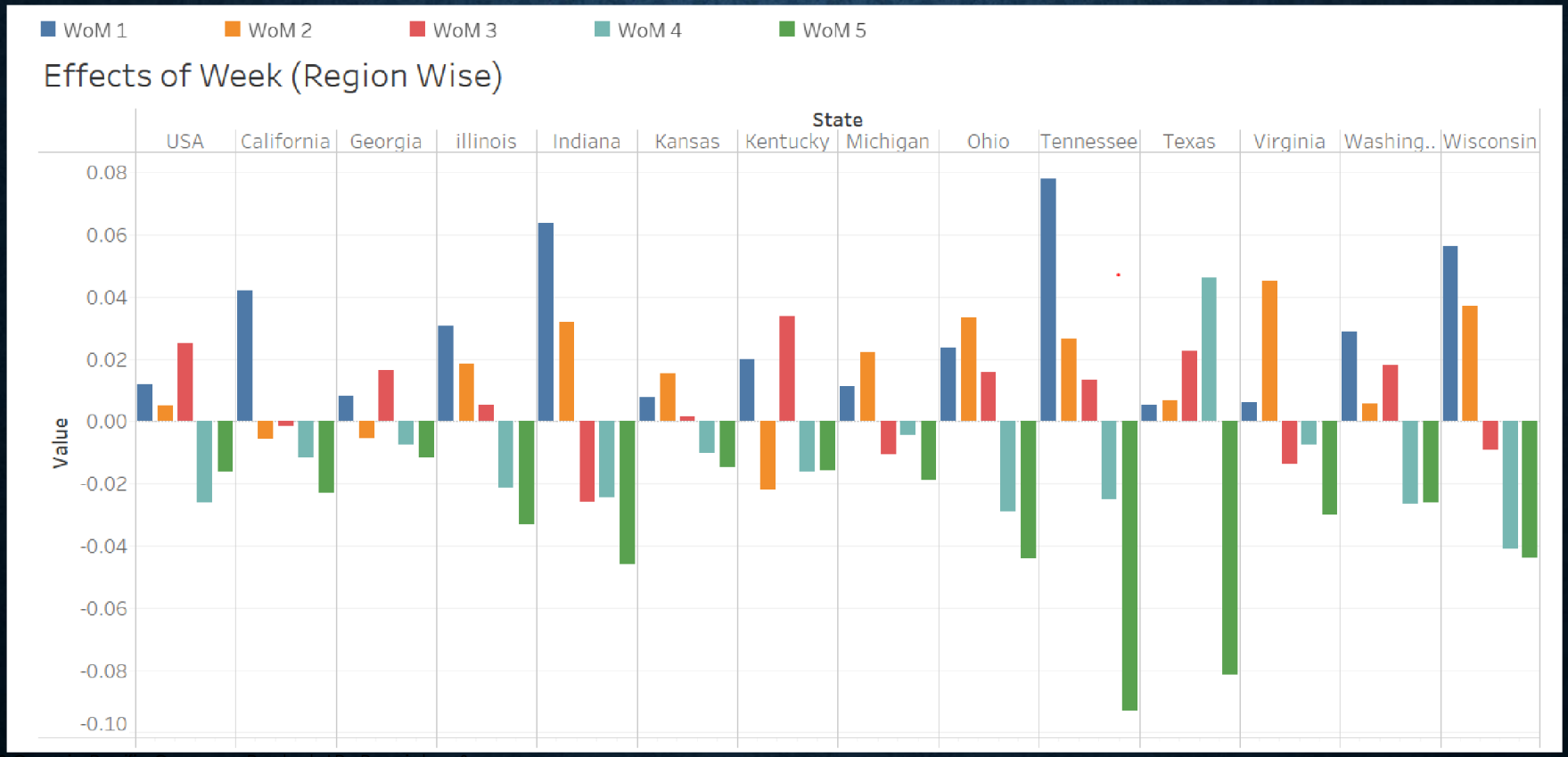
PRISM



MODEL RESULT

A clear trend is that 4 and 5 week of month do not contribute to the sell of the products. Generally, we can say that more products are sold in first and second week of the month except for the Texas region.

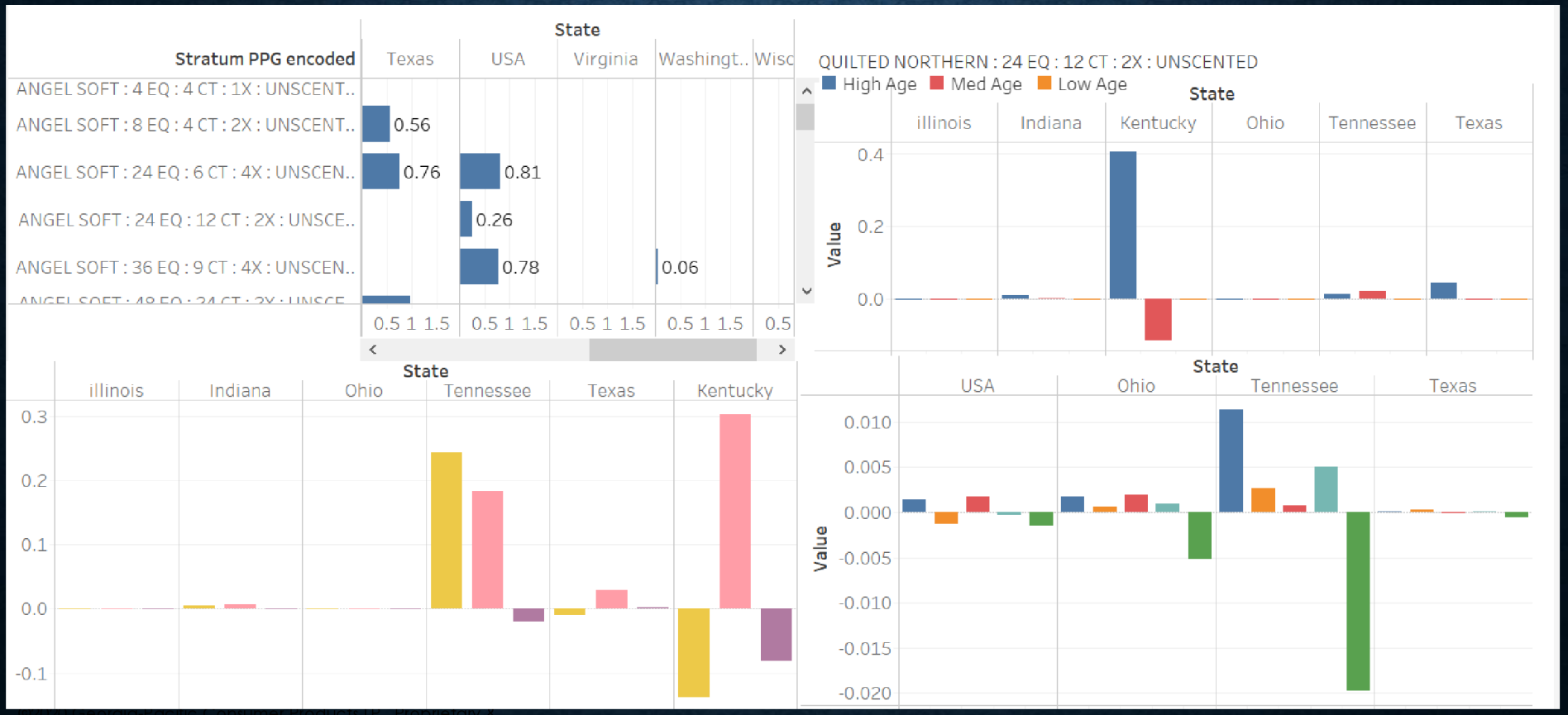
PRISM



DASHBOARD

The Dashboard helps us to select a specific Model based on the R Squared and check its performance across various region using various factors.

PRISM



USED CASE ON A PPG- ANGEL SOFT: 24 EQ: 6 CT : 4X: UNSCENTED

- Comparing the model output of different Region we can say that the PPG has display as main contributing factor in sales. Also, Display plays a relative different role in different state.
- 1 st week is more contributing to the sale in Tennessee while 3 rd week is more contributing in Ohio as well as USA as whole.
- Medium age group is generally contributing to more sales in Ohio and USA as whole but almost as no effect in Tennessee or Texas.

KEY INSIGHTS

PRISM

- Customer that are female are more likely to buy bath tissue and paper towels than male customer.
- Customer belonging to medium age group(35-65) and high age (65-75) on average bought more product than low age group people (21-35).
- Customer belonging to high income (75-100K) and medium income group (55-75K) buy more products rather than low income group (35-55 K).
- In general, the customer are more likely to buy the products in the first and second week followed by third week of a month. They are less likely to buy products in the fourth & fifth week of a month.
- For regional models, customer are more likely to buy the products on display rather than Ad (feature).
- For medium income group more sales is seen in USA, Indiana, Ohio & Virginia, while for high group more sales is seen in Georgia, Texas & Kentucky in decreasing order.
- More Product are getting sold on first week outperformed for Tennessee while second week in Virginia in comparison with other states. While over all USA , third week performed better.
- For medium age group outperformed in California state. While age group didn't really contribute in sales for Michigan state.
- For all region lowest products sale is from Wisconsin region while having moderate AD %.