# Black Friday Sales Analysis

## ABC Pvt Ltd.

by

KSHITIJ MAMGAIN


# Contents


1. Introduction
2. Analysis Objectives
3. Methodology
4. Analysis Framework
5. Descriptive Analysis
6. Hypothesis
7. Results
8. Conclusion
9. Recommendations
10. Appendix

# Introduction

- The ABC Pvt. Ltd. company wants to understand the customer purchase behavior from previous year Black Friday Sales data.
- The purchase summary of various customers for selected high volume products from last year is available.
- The available information has total 12 variables related to customer demographics, product details and purchase amount
- The data has total 550068 observations and 12 variables including the target ‘purchase’

# Analysis Objective

- The ABC company intends to analyze the data to increase business for this year by:

i) Analyzing the supply side operations – focusing on the product demand and the revenue by store locations

ii) Assessing the demand side consumer profile to develop a better target strategy

# Methodology

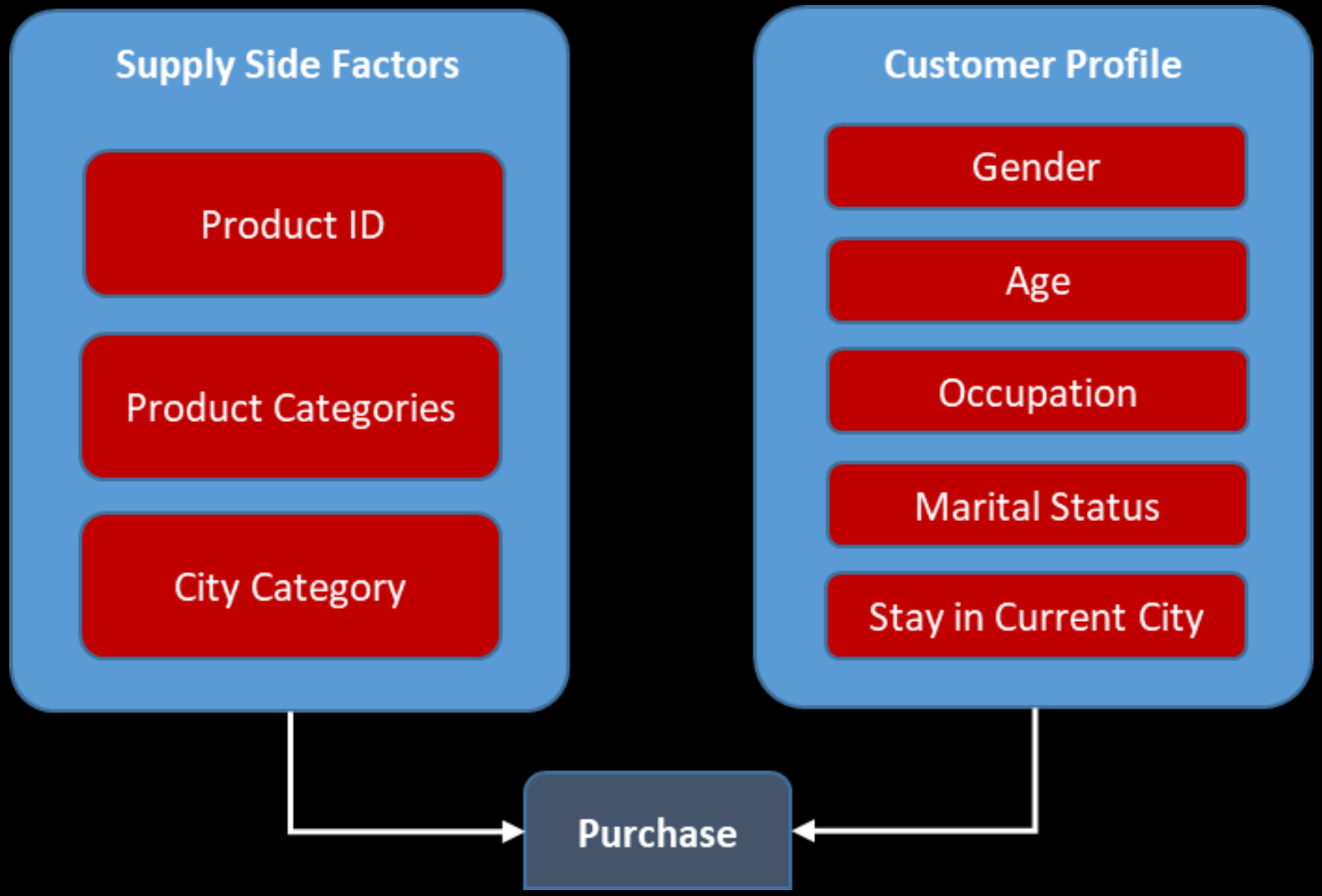
- Data Collection: .csv files from organization
- Data Validation: All the features are categorical variable while the label column purchase is continuous
- Each observation indicates a product which is classified by 3 categories
- Product Category 2 has 173638 and Product Category 3 has 383247 missing values

| Variable | Missing |
|---|---|
| Age | 0 |
| City Category | 0 |
| Gender | 0 |
| Marital Status | 0 |
| Occupation | 0 |
| Product Category_1 | 0 |
| Product Category_2 | 173638 |
| Product Category_3 | 383247 |
| Product ID | 0 |
| Purchase | 0 |
| Stay In Current City Years | 0 |
| User ID | 0 |

# Methodology (contd.)

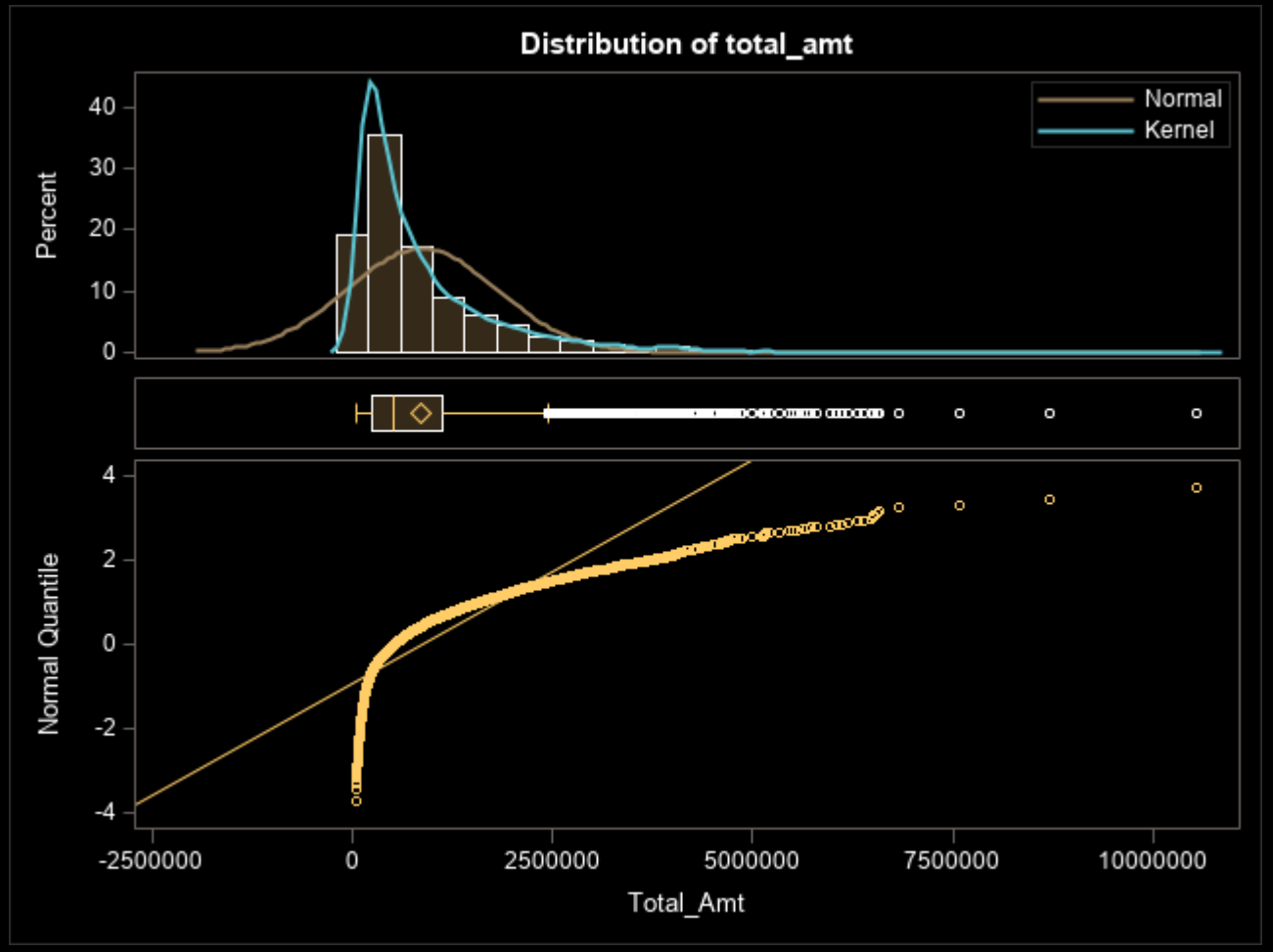
- Feature Engineering: For business case we drop the Product Category 3 feature and merge the values in Product Category 1 and Product Category 2
- Feature Transformation: The User ID and Product ID carry no general information but provides counts for customer and products which is used to transform the dataset to create unique product and cusomer count
- Similarly the new feature from product category creates a new column for category wise product distribution

# Analysis Framework



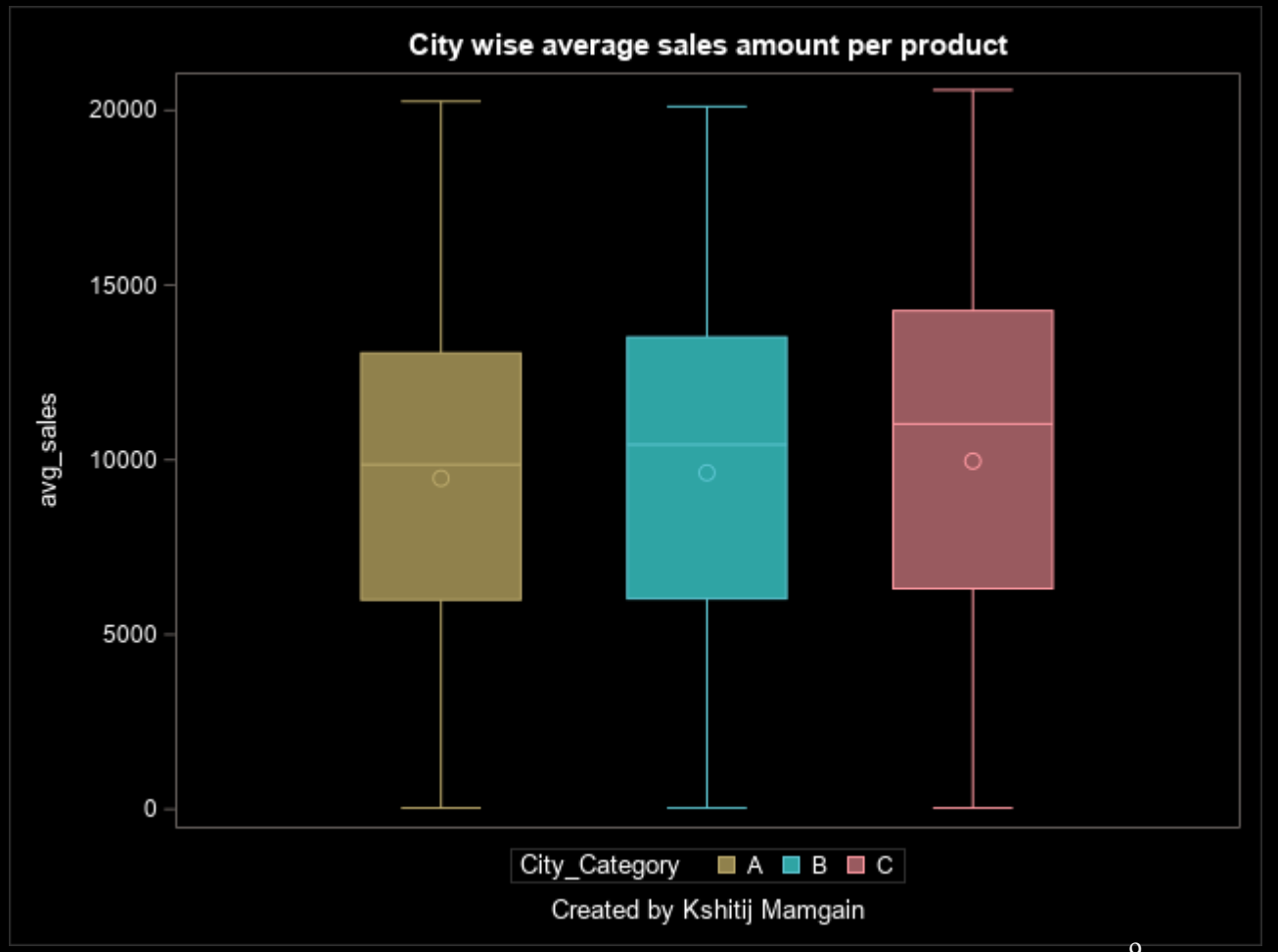
# Purchase Distribution

- (23859,11930,8014,5397,24)
- Mean: 9073.303
- Std Deviation: 4994
- Range: 23835



# Purchase Distribution

- (23859,11930,8014,5397,24)
- Mean: 9073.303
- Std Deviation: 4994
- Range: 23835



# Customer Distribution

- The dominant age group for customers is 26-35 years which together are ~40% of customers
- The number of products are purchased most in Category B type cities and least in category A cities
- Customers who are single are more likely to come for sales than married people

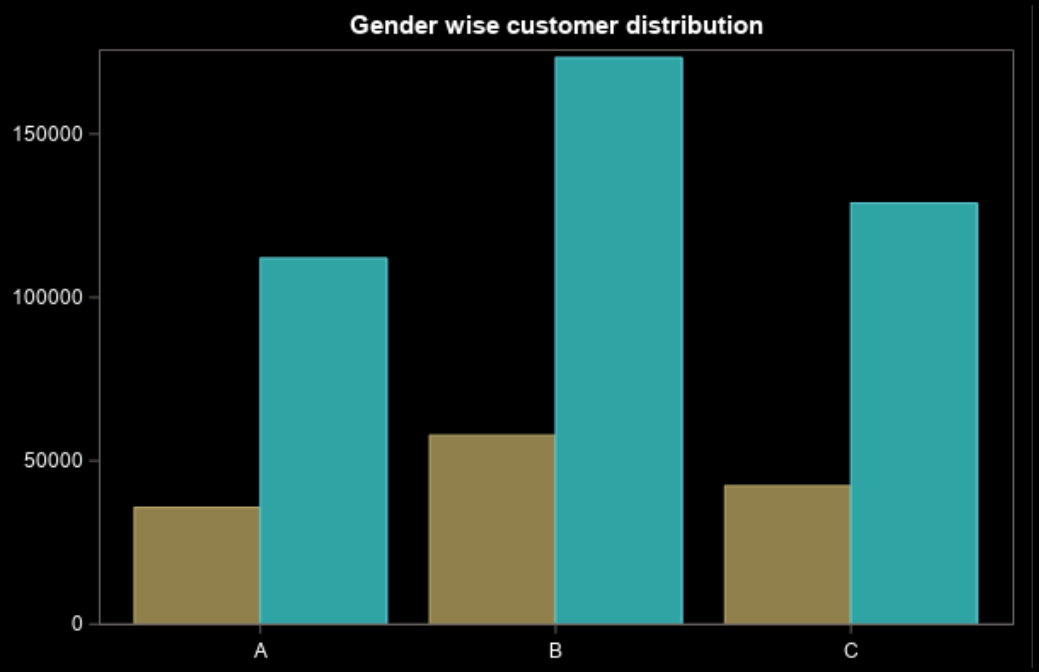
| | | Age groups | | | | | | | All | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | 0-17 | 18-25 | 26-35 | 36-45 | 46-50 | 51-55 | 55+ | | |
| | | % | % | % | % | % | % | % | N | PctN |
| City_Category | Marital_Status | | | | | | | | | |
| A | 0 | 0.46 | 3.96 | 8.43 | 2.68 | 0.44 | 0.32 | 0.28 | 91173 | 16.57 |
| | 1 | . | 1.04 | 4.98 | 2.16 | 0.94 | 0.79 | 0.37 | 56547 | 10.28 |
| | All | 0.46 | 5.01 | 13.41 | 4.84 | 1.38 | 1.11 | 0.65 | 147720 | 26.85 |
| B | Marital_Status | | | | | | | | | |
| | 0 | 0.99 | 6.12 | 10.22 | 5.48 | 0.88 | 0.87 | 0.30 | 136721 | 24.86 |
| | 1 | . | 1.74 | 6.43 | 3.18 | 2.83 | 2.36 | 0.63 | 94452 | 17.17 |
| | All | 0.99 | 7.86 | 16.65 | 8.65 | 3.71 | 3.23 | 0.94 | 231173 | 42.03 |
| C | Marital_Status | | | | | | | | | |
| | 0 | 1.29 | 4.20 | 5.59 | 3.91 | 0.98 | 0.78 | 0.85 | 96837 | 17.60 |
| | 1 | . | 1.05 | 4.28 | 2.59 | 2.23 | 1.89 | 1.47 | 74338 | 13.51 |
| | All | 1.29 | 5.25 | 9.86 | 6.51 | 3.22 | 2.67 | 2.32 | 171175 | 31.12 |

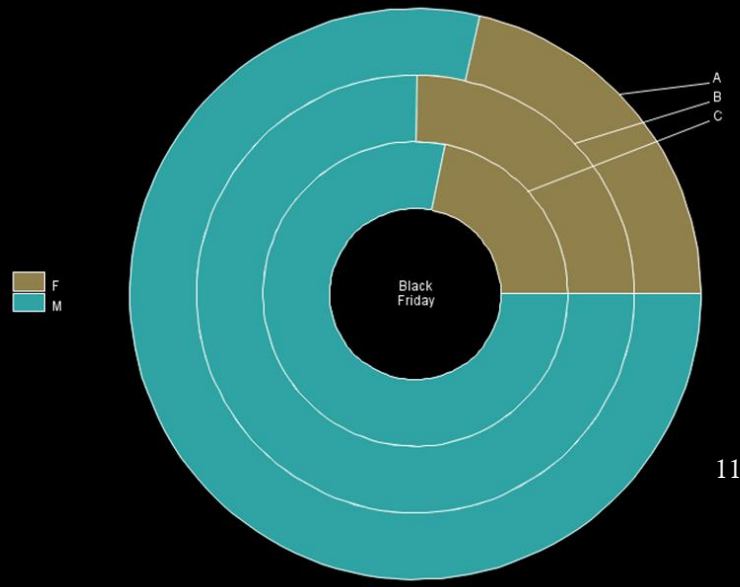
# Customer Distribution

- Contrary to belief Men are more likely to shop in organized sales as is visible across all the city types

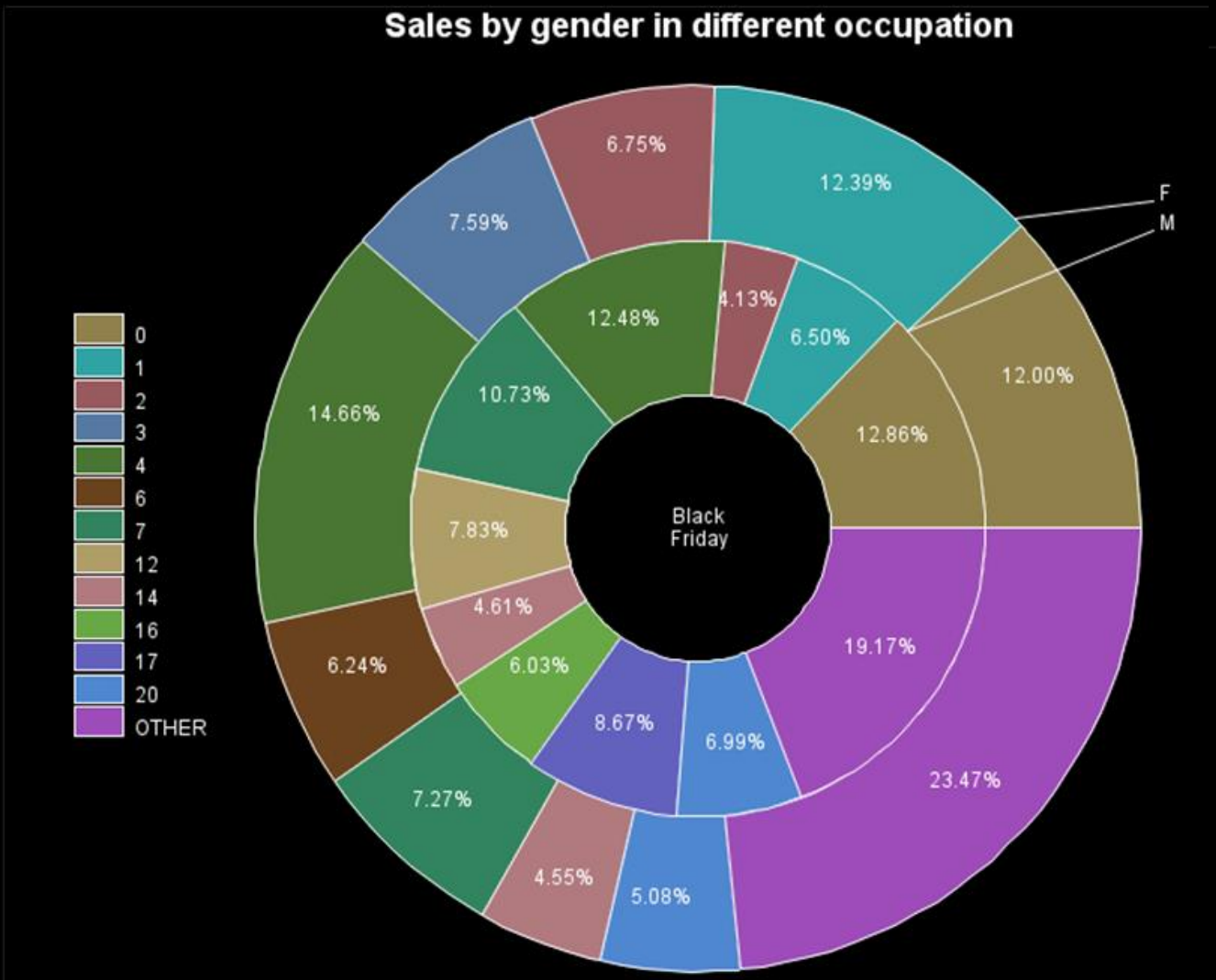
**Customer count in city categories**

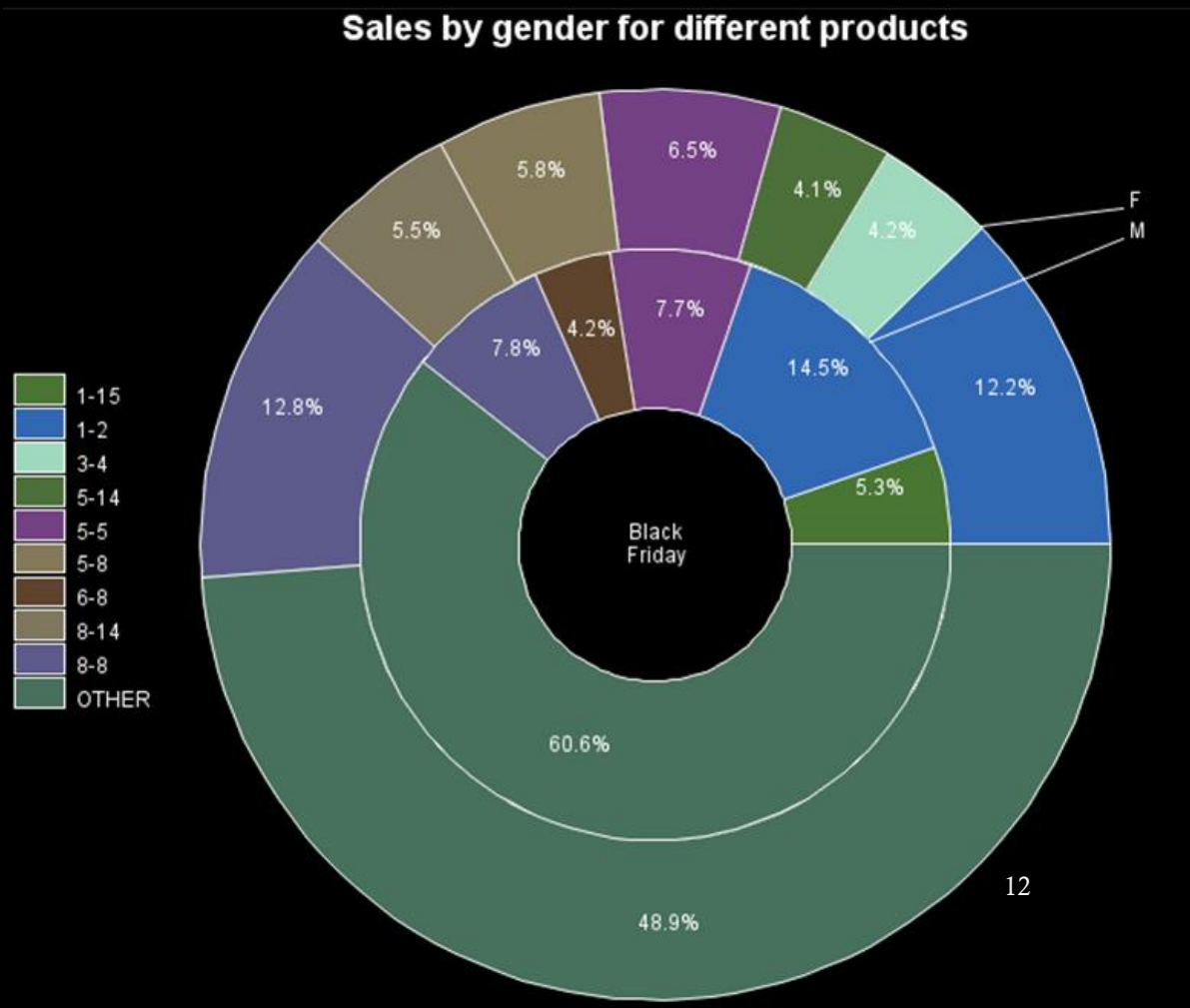
| | Gender | |
|---|---|---|
| | F | M |
| | Sum | Sum |
| City_Category | | |
| A | 295.00 | 750.00 |
| B | 503.00 | 1204.00 |
| C | 868.00 | 2271.00 |
| All | 1666.00 | 4225.00 |



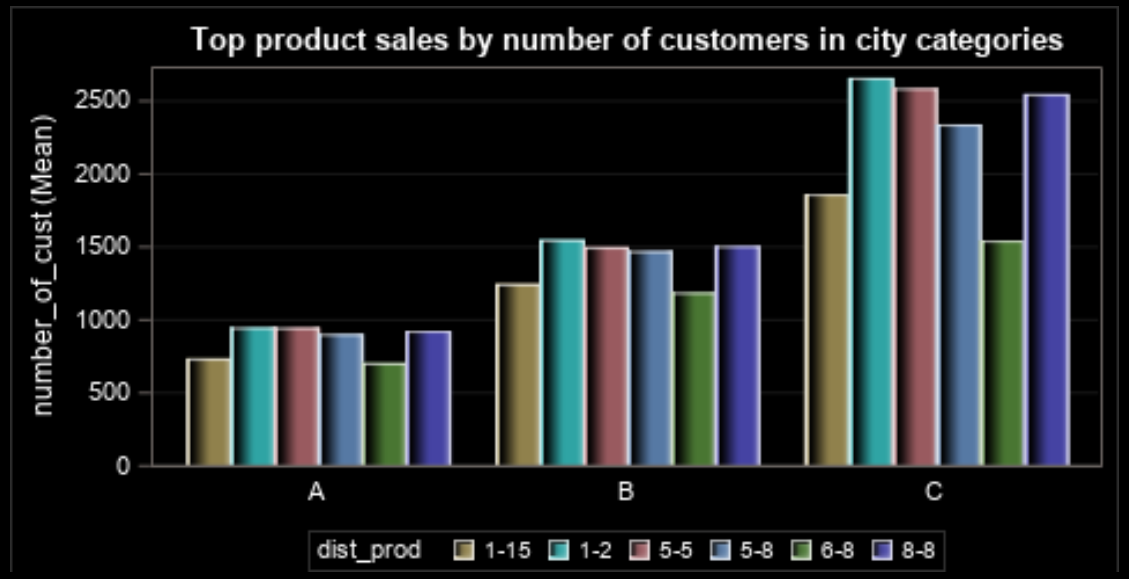
# Gender-wise Analysis

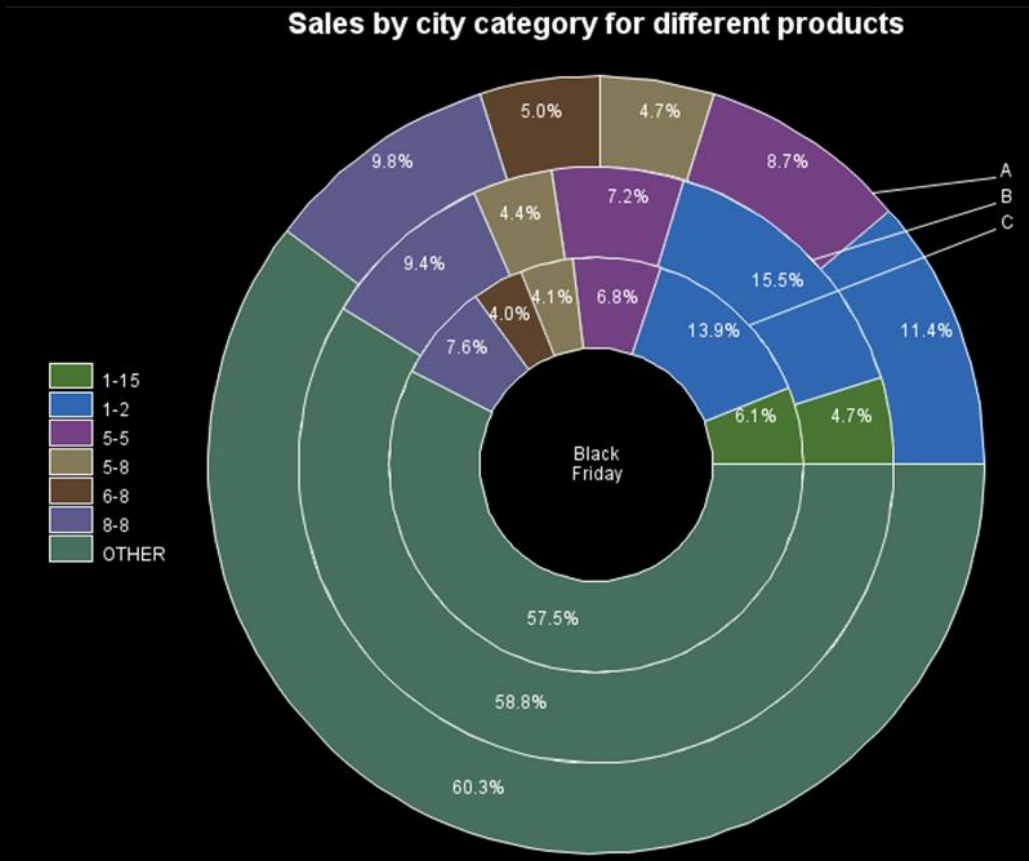


# City category wise Analysis

- A look at distribution of product being purchased across different cities, shows that the demand is alike
- 40% of total sales have come from basket of 6 distinct product categories

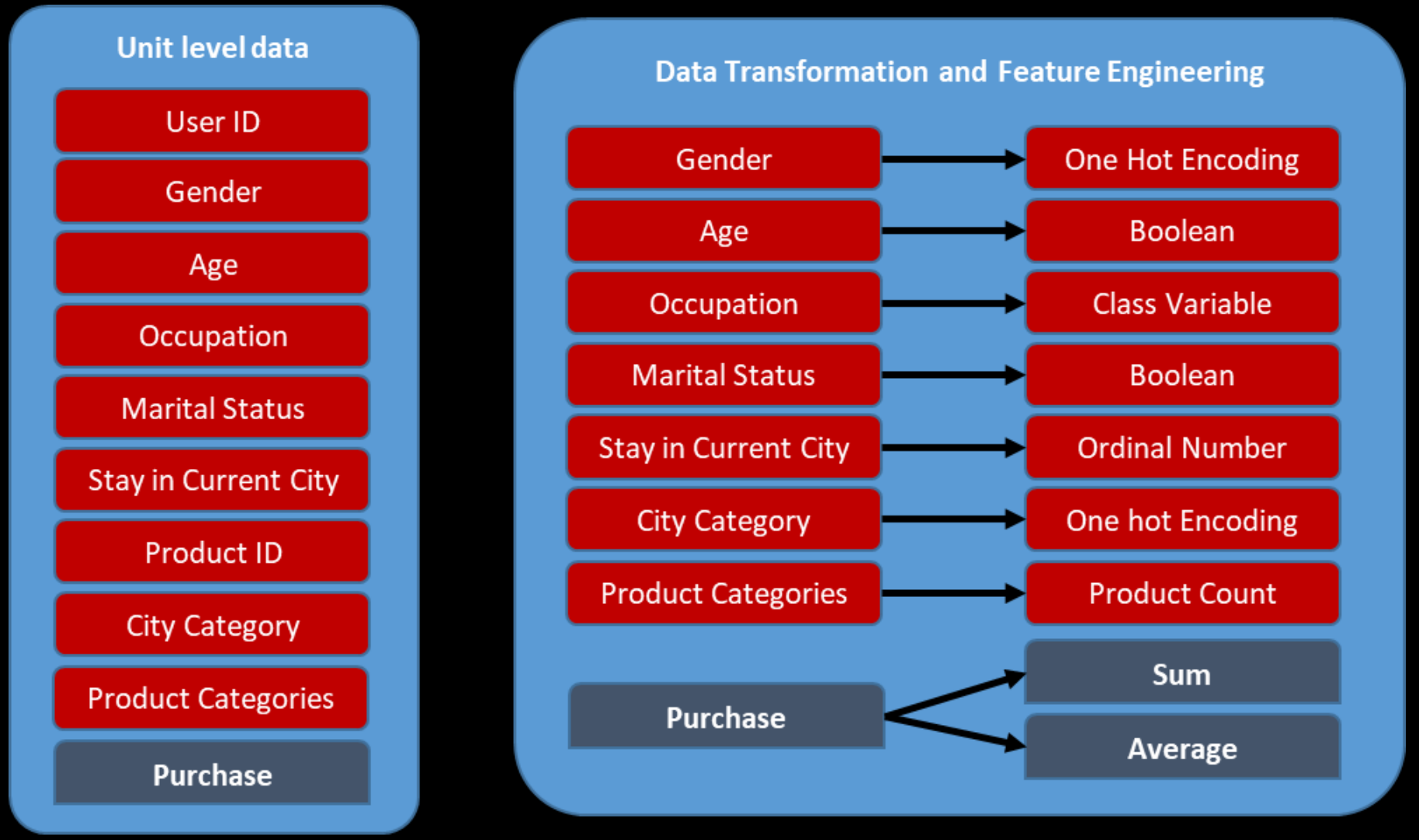


# Research Questions

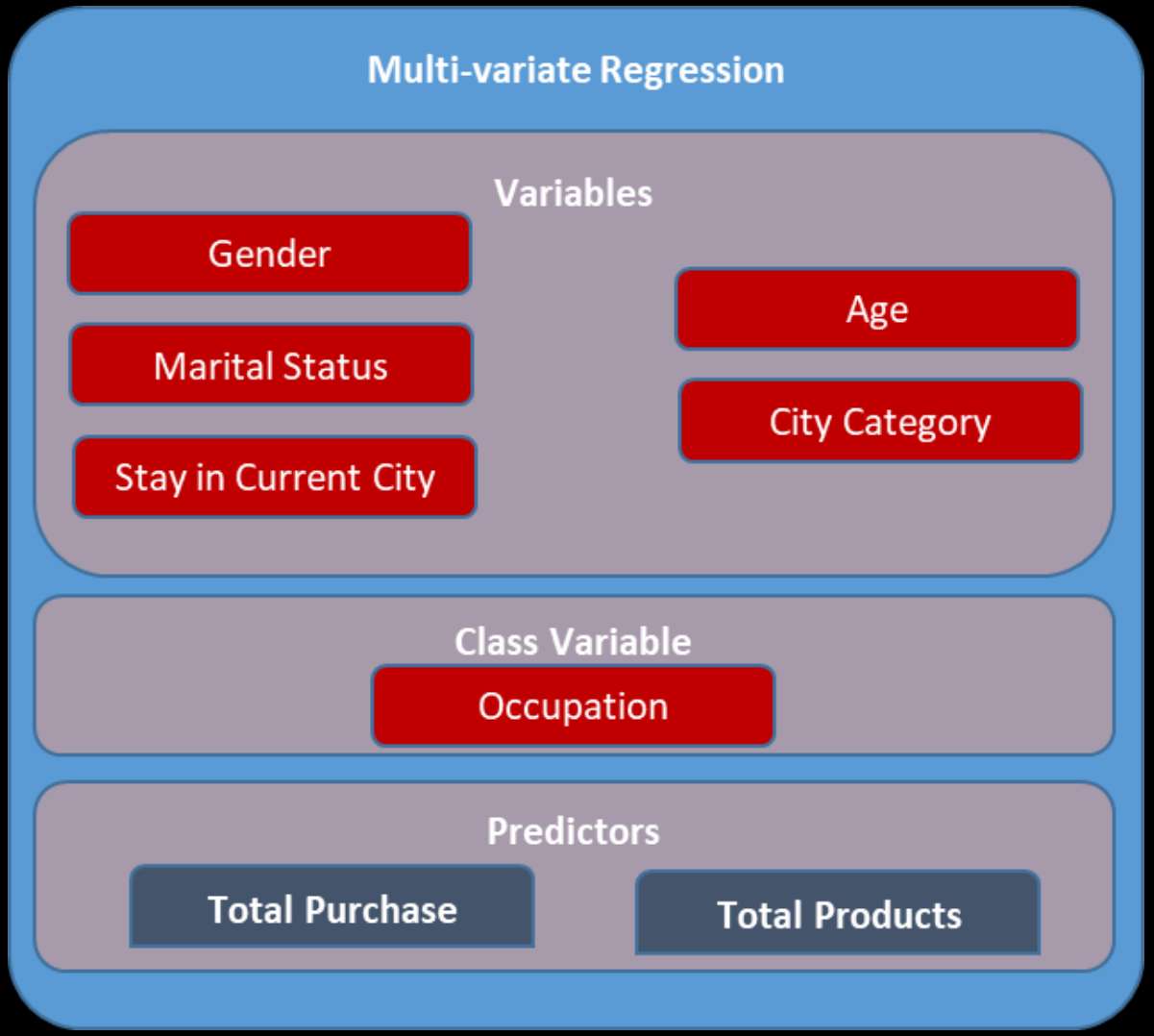
- Which variables correlate with the purchase?
- Which customer profile are more likely to make sales purchases?
- How well can the purchase be predicted from correlating the variables?

# Data Preparation & Feature Transformation



# Data Modelling

- From different procedure for Regression GLM is best suited since the dataset has mainly Categorical Data
- Occupation feature was treated as Class variable while rest were treated as continuous
- All feature were standardized with mean 0 and variance 1
- A preliminary Regression analysis also helps to understand the feature importance



# Hypothesis Testing and Summary Statistics

Running Anova, Chi Square and Correlation procedures for establishing the relationships with multiple features.

| Chi Square test for features | | P>r |
|---|---|---|
| Age | Gender | 0.0023 |
| Occupation | Gender | <.0001 |
| Age | Occupation | <.0001 |

| Product-wise Count (Corr) | P>r |
|---|---|
| Total Sales | <.0001 |
| Average Sales | 0.5999 |

| Customer Purchase (Anova) | Pr>F |
|---|---|
| Gender | <.0001 |
| City Category | <.0001 |
| Marital Status | 0.1370 |
| Occupation | <.0001 |
| Age | <.0001 |
| Stay In Current City Years | 0.4426 |

| Distinct Product Sales (Anova) | Pr>F |
|---|---|
| Number of purchases | <.0001 |
| Average Sales | <.0001 |
| City Category | <.0001 |

# Feature Selection for Prediction

- We've seen from the hypothesis testing that customer Age, Gender, Occupation and Stay in Current City have a correlation with number of products purchased and the sales.
- City Category and the Product Type also have an impact on sales

**Test with GLM Procedure**

| Parameter | Estimate | Standard Error | t Value | Pr > \|t\| |
|---|---|---|---|---|
| Intercept | 865016.5918 | 11298.55387 | 76.56 | <.0001 |
| Sex | -100992.2465 | 11302.60635 | -8.94 | <.0001 |
| City Category A | 276754.3031 | 11887.57826 | 23.28 | <.0001 |
| City Category B | 321493.3092 | 11874.27630 | 27.07 | <.0001 |
| Median Age | -25022.2578 | 11371.61640 | -2.20 | 0.0278 |

Parameters
Estimate insight

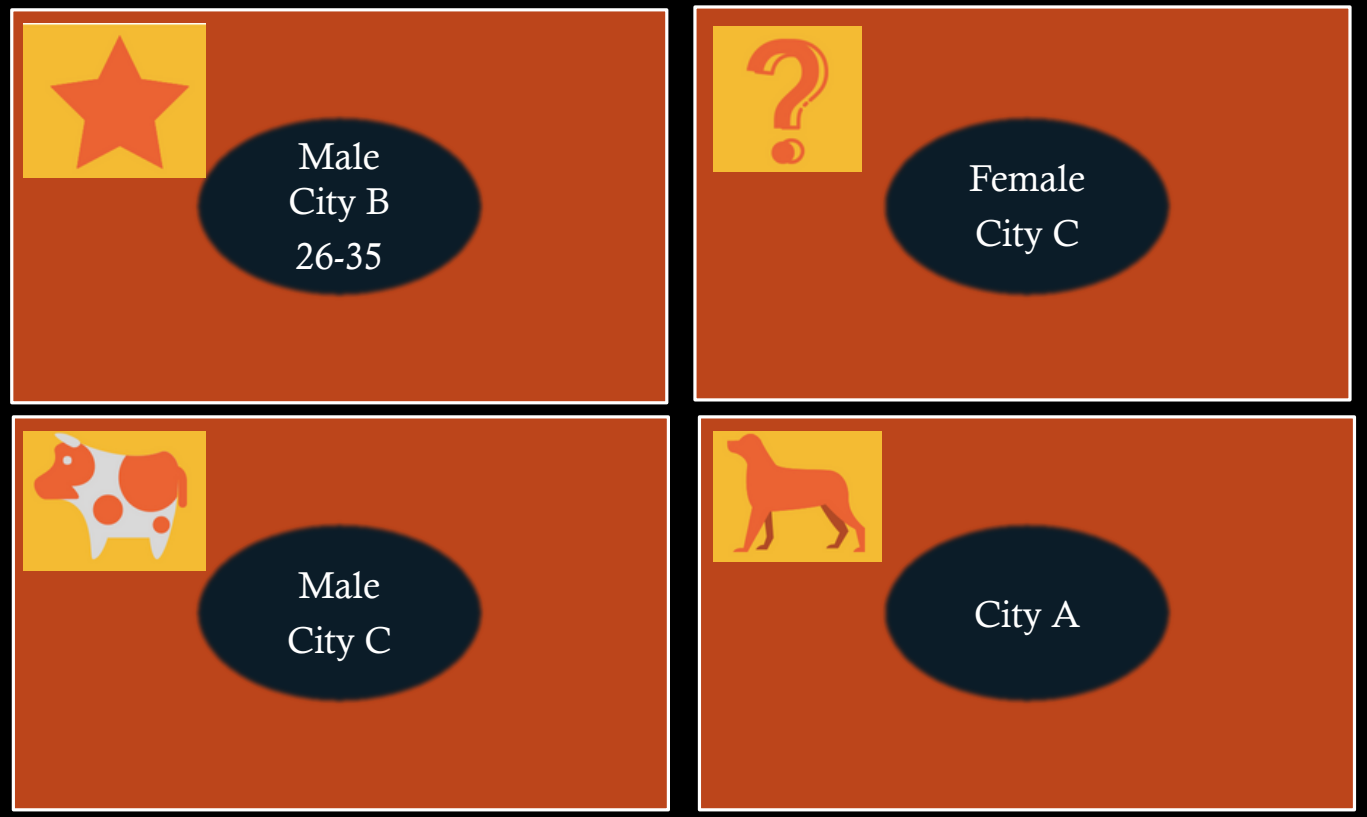
Low $R^2$
Not suitable for prediction

# Customer Segmentation on Significant Features

| | | Age groups | | | | | | | All | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | 0-17 | 18-25 | 26-35 | 36-45 | 46-50 | 51-55 | 55+ | | |
| | | % | % | % | % | % | % | % | N | Pct N |
| City Category | Gender | | | | | | | | | |
| A | F | 0.24 | 0.95 | 2.09 | 0.90 | 0.31 | 0.36 | 0.17 | 295 | 5.01 |
| | M | 0.19 | 2.68 | 5.74 | 2.09 | 0.59 | 0.78 | 0.66 | 750 | 12.73 |
| | All | 0.42 | 3.63 | 7.83 | 2.99 | 0.90 | 1.14 | 0.83 | 1045 | 17.74 |
| B | Gender | | | | | | | | | |
| | F | 0.31 | 1.60 | 3.12 | 1.66 | 0.98 | 0.61 | 0.25 | 503 | 8.54 |
| | M | 0.54 | 4.02 | 7.94 | 4.02 | 1.49 | 1.68 | 0.73 | 1204 | 20.44 |
| | All | 0.85 | 5.62 | 11.07 | 5.69 | 2.48 | 2.29 | 0.98 | 1707 | 28.98 |
| C | Gender | | | | | | | | | |
| | F | 0.78 | 2.33 | 4.04 | 3.09 | 1.80 | 1.44 | 1.26 | 868 | 14.73 |
| | M | 1.65 | 6.57 | 11.92 | 8.05 | 3.84 | 3.29 | 3.24 | 2271 | 38.55 |
| | All | 2.43 | 8.89 | 15.96 | 11.14 | 5.64 | 4.74 | 4.50 | 3139 | 53.28 |

# Market Strategy

- Cow: Male in City C contributed to around 40% of revenue
- Star: Males in City B contribute to 20% or overall sales
- Question Mark: Female in City C can be part of exploration strategy to increase the purchase s it has high count
- Dog: Overall City A lags behind and would need overhaul



# Analysis Findings

- The city category, customer age group and gender are the key influencer in increasing purchases

1. Majority of purchases were made by customers in city category C.
2. Majority of purchases were made by customers in age group of 26-35 .
3. Majority of purchases were made by men.

- Marital status and stay in the city had no statistically significant impact on the sales purchases

# Conclusion

- The analyzed variables have a significant impact on the total sales and next sales strategy can focus on the segmented customers to increase sales.
- The sales team can choose an appropriate strategy:

1. To increase the current market stronghold ‘Star’ by focusing on men customers in City Category B
2. To innovate and target women  customers in City Category C and explore new avenues for customer outreach
3. Assess the short coming and poor performance in City Category A

# Assumptions and Shortcomings

- The dataset is available for a single day sales in past year therefore current analysis is unable to comment on trend and the prospective growth in the sales and revenue
- The regression analysis presented a very low R2 score which points to further study on intermittent and all available data to draw more insights on the customer behaviour and overall trend and seasonality in the sales


# Thank You

# Appendix

## Procedures Used

- PROC CONTENTS
- PROC FREQ
- PROQ MEANS
- PROC UNIVARIATE
- PROC SQL
- PROC SGPLOT
- PROC ANOVA
- PROC IMPORT
- PROC TEMPLATE
- PROC PRINT
- PROQ SORT
- PROC TABULATE
- PROC GCHART
- PROC CORR
- PROC GLM
- PROC REG
- PROC FORMAT