

mis771 – DESCRIPTIVE ANALYTICS AND VISUALISATION REPORT



NAME: SWASTIK AIREE STUDENT ID: 224389999

INTRODUCTION

The function of a business analyst is crucial and has the power to make or break a firm. Here, we'll look at Cindy Varanasi, a Senior Partner at Method9, a business that provides analytical solutions to businesses. Making data models for different company tasks is my role. We have received the unique analytical requirements from each organization, which we will address in detail in this report. Each company's analytical requirements are broken down into discrete tasks, which are covered in more detail in the report's main body.

ANALYSIS

TASK 1:

**GOAL**: To create a predictive model for Grocery Plus sales forecasting by applying suitable diagnostic techniques and iterations.

**ACTIONS TAKEN:**

* The variables "Wages," "Number of Staff," "Advertisement Expenditure," and "Car Spaces" showed strong association coefficients (> 0.8) after the correlation matrix was analyzed. Strong linearity was confirmed by the scatter plots.
* Given the strong correlation (0.92) seen between "Wages" and "Number of Staff," multicollinearity was addressed by removing the variable "Number of Staff." "Car Spaces" (p-value = 0.67) was the first unimportant variable to be eliminated using the backward elimination procedure.
* To narrow the results further, non-significant variables were eliminated one at a time in subsequent iterations.

**RESIDUAL ANALYSIS:**

Tests were conducted in response to indications of potential normalcy violations in the residual diagnostics. It was believed that employing a Box-Cox transformation could rectify the issue and enhance residual normality.

Final Model Equation: Sales = 1.73+2.43\*Wages+0.02\*Adv.$'000Final Model Equation: Sales = 1.73+2.43\*Wages+0.02\*Adv.$'000

TASK 2:

**GOAL:** Bike Mart has assigned us the duty of researching the effects of advertising campaigns and marketing expenditures on sales, especially when they combine.

**ACTIONS TAKEN:**

* We created a new variable called "Advertisement \* Promotions" to account for the interaction effect between the quantity of promotional campaigns and advertising budget.
* Regression analysis was employed to evaluate the respective and collective impacts of these variables on sales.

**Presumptions:**

* **Linearity:** The connection between the independent variables (advertising and promotions) and the dependent variable (sales) is assumed to be linear.
* **Independence:** The observations don't correlate with one another.
* **Normality:** There is a normal distribution among the residuals of the model.
* **Homoscedasticity:** At every level of the independent variables, the residuals' variance stays constant.
* The independent variables and the interaction term do not exhibit complete multicollinearity.

**OBSERVED RESULTS:**

Two-Way ANOVA: The interaction term and advertising spending demonstrated a statistically significant impact on sales (p-values < 0.05), indicating a positive association. However, promotions alone had minimal effect on sales.

Interaction Effect: The visual plot indicates an interaction effect due to the non-parallel lines. The number of promotional campaigns alters the relationship between advertising expenditure and sales, exerting a stronger influence when both variables are high.

**RESIDUAL ANALYSIS:**

A residual analysis was done to make sure the assumptions were upheld. The model's validity was confirmed when the residuals were compared to the fitted values and no notable breaches were found.

A graph with lines and points

Description automatically generated with medium confidence

TASK 3:

**GOAL:** Creating a prediction model for Gadget4U required figuring out how likely it was that a customer would buy headphones after making a smartphone purchase.

**MEASURES TAKEN:**

* Using a correlation matrix, a number of independent variables, including age, gender, click-through rate on advertisements, annual income, past purchases, days since the last purchase, and click-through rate on advertisements, were compared to the dependent variable (headphone purchase) to determine their relevance and potential multicollinearity issues.
* By employing a backward elimination technique, inconsequential variables were gradually removed in order to guarantee that the final model had only the most pertinent variables. Both practically and statistically significant variables, including age and gender, were included in the final logistic regression model.

**RESULTS OBSERVED:**

* With a 74.66% prediction accuracy, the model outperformed the PCC Hit Ratio (59.10%) and the Standard Hit Ratio (73.87%).
* AUC (Area Under the Curve) of 73.52% indicates a great discriminatory ability of the model.
* According to analysis, the likelihood of buying headphones increased by 72% and 118%, respectively, depending on gender and clicked advertisements.

TASK 4.1

**GOAL:** Using gender, age, and experience as determinants, a logistic regression model will be created for the Cosmetic Chain to predict the chance of a shop manager quitting.

ACTIONS MADE:

1. **Dummy encoding:**

* Gender: 1 for men and 0 for women.
* Status of Resignation: Yes = 1, No = 0.

1. **Logistic Regression Model:** To predict the chance of resigning, the model was built utilizing age, gender, and experience. After converting the categorical data into dummy codes, logistic regression was applied.
2. **Model Execution:**

* **Precision:** The model outperformed the proportion of correct classification (PCC) hit ratio (52%) and the standard hit ratio (64%), achieving an accuracy rating of 77%.
* **The AUC** (Area Under the Curve) of 86% indicates a good ability to distinguish between people who are resigning and those who are not.
* **Chi-Square:** A positive chi-square value indicates that the model outperforms a basic baseline model by a significant margin.
* **Pseudo R-Square:** As is common with logistic regression models, it shows a moderate level of explanatory power.

**Interpretation of Odds Ratio:**

* Experience: The likelihood of quitting increases by 46% for every extra year of experience.
* Gender: Managers who are male have a 164% higher resignation rate than those who are female.
* Age: There is a modest negative influence of age on resignation likelihood, suggesting that older managers are marginally less inclined to quit.

**The resulting regression model is shown below:**

Resigned Dummy = 0.42 + (-0.13) \* Age Manager + 0.38 \* Experience Manager + 0.97 \* Gender Manager Dummy is the regression model.

TASK 4.2

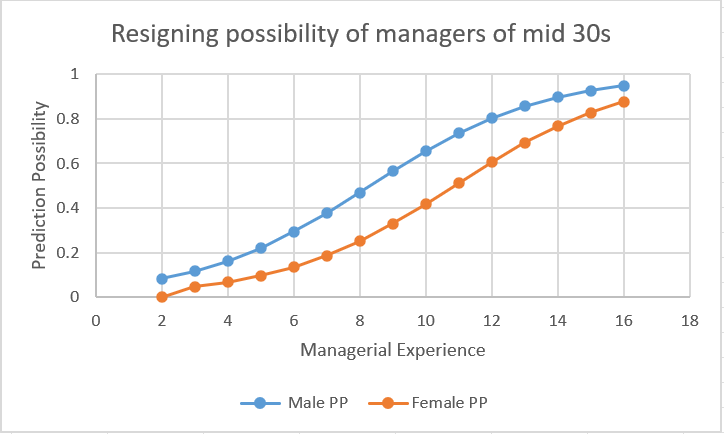
**GOAL:** Based on the shop manager's age (mid-30s), experience level (two to sixteen years), and gender, visualize and assess the likelihood that they will leave.

**STEPS TAKEN:** To investigate the relationships between these variables and the probability of resignation, we employed a logistic regression model.

* We used this formula to determine the odds: ODDS equals e^(0.42 + (-0.13) \* Gender + 0.97 \* Experience + Age).
* Next, we divided the equation to determine the odds for men and women: Odds With ODDS, Male is equal to e^(0.42 + (-0.13) \* Age + 0.38 \* Experience + 0.97 \* Male (coded as 1)) E^(0.42 + (-0.13) \* Age + 0.38 \* Experience + 0.97 \* Female (coded as 0)) is the female component.
* We utilized the following formula to get the Predicted Possibility (PP): PP = e^(0.42 + (-0.13) \* Gender, Age, Experience, and 0.97 \* \* Experience) / (1 + e^(0.42 + (-0.13) \* Gender, Experience, and 0.97 \* Gender)).
* This allowed us to calculate the expected possibility values for managers who were in their mid-thirties and had between two and sixteen years of experience.

**RESULTS:**

* As managerial experience increased, so did the likelihood of quitting.
* At all experience levels, male managers were more likely to resign than female managers.
* Male resignation rates for managers with at least two years of experience were 8%, while female resignation rates were 3%.
* When managers attained 16 years of experience, the likelihood of their resigning was 87% for women and 95% for men.



TASK 5:

**GOAL:** The objective was to develop a time series model that would forecast pale ale production levels over the course of the following four fiscal quarters.

**MEASURES TAKEN:**

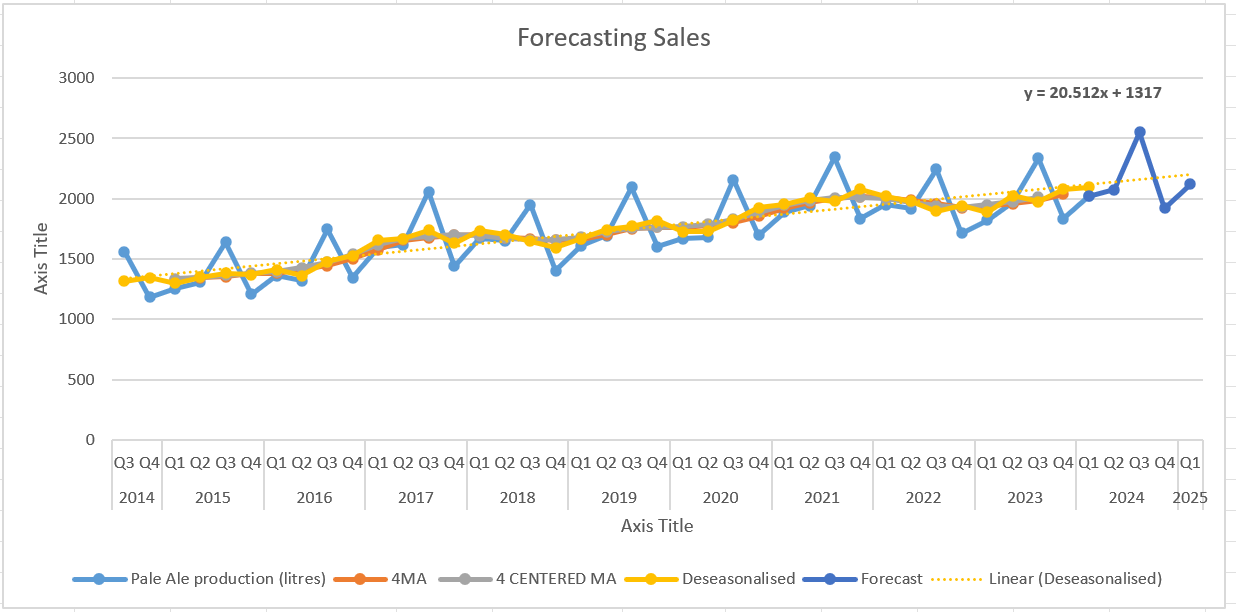
* Production quantities (in litters) for every quarter of every year were included in the dataset, except for 2014, when only Q3 and Q4 data were available.
* Quarters were labelled in chronological order, beginning with Q3 of 2014 and ending with Q1 of 2024.
* Pale ale production generally climbed from Q1 to Q3, peaking in Q3 and then declining in Q4. But from Q3 of 2014 to Q2 of 2024, the total production capacity trended upward, suggesting that the data followed a seasonal pattern.
* The data series was smoothed using the moving averages approach in order to eliminate the seasonal volatility. Four data points at a time were averaged using a 4-period moving average, which was applied considering the seasonal pattern with four separate behaviours throughout quarters.
* The time series chart was then displayed with a 4-cent moving average. Since time series often entail these components being multiplied together, a multiplicative model was employed for forecasting because the data contained seasonal, trend, and irregular components.
* The observed values were divided by the moving averages in order to get the trend component. Next, quarter tables and the following seasonal index equation were used to extract the seasonal effect:

Indices are equal to (4 / Sum of Averages) \* (Q1 / Q2 / Q3 / Q4).

* Based on the seasonal pattern, Q3's production levels were the highest. The time series was separated by the seasonal component in order to exclude the seasonal volatility from the data.
* The time series chart was then updated with the deseasonalized data.
* The deseasonalized data were fitted with a trend-based model, yielding the following model equation:
* y = 20.512x + 1317, R² = 0.9026.
* The deseasonalized trend values for the following four quarters were calculated using this equation and are indicated in yellow. However, more modifications were required because seasonality was not considered in these numbers.
* To increase the forecast's accuracy, the deseasonalized trend values were re-seasonalized.
* The final re-seasonalized values, with the seasonal component added back in, were forecasted for the next four quarters and highlighted in blue on the time series chart.

**RESULTS OBSERVED:**

\* Absolute Percentage Error (APE) was utilized to evaluate accuracy in the absence of actual data for the next four quarters. A 96.3381% forecast accuracy was suggested by the average APE of 3.6619%.



**CONCLUSION**

In summary, we effectively met the analytical objectives of multiple organizations—from forecasting sales to estimating manufacturing volumes—through careful investigation and model building. Even with multicollinearity and statistical presumptions, our models have produced insightful analyses and precise predictions. These businesses have benefited from our data, which have allowed them to make wise choices about marketing tactics, customer purchases, and staff churn. These models need to be improved and altered in order to continue to be useful in changing business environments.