# **Analysing Sales Data for a Retail Store**

## **Introduction**

* Briefly explain the purpose of the project and the data being analysed
* Describe the dataset and its features (e.g., format, size, structure, etc.)
* State the research questions/hypotheses that will guide the analysis

## **Data Pre-processing**

* Describe the steps taken to clean and prepare the data for analysis
* Highlight any missing data, outliers, or other anomalies and explain how they were dealt with
* Discuss any transformations or feature engineering that was done to the data

## **Exploratory Data Analysis**

* Present summary statistics and visualizations of the data
* Identify any trends, patterns, or relationships in the data
* Test any hypotheses using appropriate statistical methods

## **Feature Selection and Engineering**

* Determine which features are most important for predicting the outcome variable
* Engineer new features that might improve the predictive power of the model
* Discuss any challenges or limitations encountered during this process

## **Model Building and Evaluation**

* Train and evaluate several models using appropriate performance metrics
* Compare and contrast the performance of each model
* Discuss the trade-offs between model complexity and predictive power

## **Conclusion**

* Summarize the key findings and insights from the analysis
* Highlight any limitations of the study and potential avenues for future research
* Discuss the practical implications of the study and how the results can be used to inform decision-making

### **Example**

For example, let's say we want to analyse sales data for a retail store. Our dataset includes information on customer demographics (age, gender, income, etc.), purchase history (products bought, quantity, price, etc.), and store location (region, city, etc.). Our research question is: "How can we optimize sales and marketing efforts to increase revenue?"

After cleaning and preparing the data, we perform exploratory data analysis and find that sales are highest among young adults with moderate incomes, who tend to buy a mix of high-priced and low-priced items. We also find that sales vary significantly by region, with the highest sales occurring in urban areas.

Based on these findings, we decide to engineer a new feature that captures the "price sensitivity" of each customer, as well as a feature that indicates whether a customer lives in an urban area or not. We then train and evaluate several models, including linear regression, decision trees, and random forests, and find that the random forest model performs best in terms of predictive power.

Finally, we conclude that the key to increasing revenue is to target young adults with moderate incomes who live in urban areas, and to focus on promoting a mix of high-priced and low-priced items. We also note that future research could explore other factors that influence sales, such as seasonal trends, store layout, and marketing campaigns.