**Iowa Liquor Sales 2020-2021 Sales**



**Step 1: Define the business Problem:**

**Business Problem**:  
The goal of this project is to predict liquor sales in Iowa to optimize inventory management, improve promotional strategies, and understand consumer purchasing patterns. By analyzing historical sales data, the model will predict future sales, identify trends, and help optimize stock levels for different liquor products across various store locations.

**Potential Research Questions**:

* How can we predict liquor sales for a specific store location based on historical data?
* What factors (e.g., holidays, store location, pricing, promotions) most significantly influence liquor sales?
* Can we predict seasonal trends in liquor sales (e.g., more sales during the holidays or weekends)?
* How can we identify underperforming products or stores using predictive analytics?
* Did Covid influence any liquor sales?

**Step 2: Datasets:** Where are you getting your data? Describe the data that you will use to solve the problem

For the Iowa liquor sales dataset, there’s a publicly available dataset from data.gov. This dataset contains the spirits purchase information of Iowa Class “E” liquor licenses by product and date of purchase from January 1, 2020, to December 31, 2022. The dataset can be used to analyze total spirits sales in Iowa of individual products at the store level.

1. Iowa Liquor Sales dataset

* **Description**: This dataset contains transaction-level information about liquor sales in Iowa. It includes columns such as product name, sale date, quantity sold, price, store ID, and more.
* **Key Features**:
  + Date: Date of the sale.
  + Store ID: The unique identifier for each liquor store.
  + City: The city in which the store is located.
  + Zip Code: The zip code of the store.
  + Category Name: The category of the liquor (e.g., beer, wine, spirits).
  + Sale Amount: The total dollar value of the sale.
  + Sale Quantity: The number of items sold.
  + Volume Sold: The volume of liquor sold (usually in liters or gallons).
* **Access**:
  + You can download the dataset from Kaggle: Iowa Liquor Sales.
  + Or you can access the dataset directly from the Iowa Liquor Control Board’s website: [Iowa Liquor Control Board Dataset](https://www.iowa.gov/). Or from data.gov website: <https://catalog.data.gov/dataset?q=iowa+liquor+sales>

**Step 3: Methods:**

1. **Data Preprocessing**:
   * **Data Cleaning**: Handle missing values, remove duplicates, and ensure all dates are in a consistent format.
   * **Feature Engineering**: Extract useful features such as the day of the week, month, or season from the sale date. You could also create features based on weather or holidays if using supplementary data.
2. **Exploratory Data Analysis (EDA)**:
   * Visualize trends in liquor sales over time (e.g., monthly, weekly).
   * Examine sales by store, category, and location.
   * Identify outliers, seasonal effects, and trends using time-series plots.
3. **Time Series Analysis**:
   * **ARIMA (AutoRegressive Integrated Moving Average)**: ARIMA is commonly used for forecasting time-series data. It can help predict future liquor sales based on past sales trends.
   * **Exponential Smoothing**: A method that gives more weight to recent observations, often useful in demand forecasting.
   * **Facebook Prophet**: A flexible forecasting tool that works well with seasonality and holiday effects, which can be helpful for modeling seasonal liquor sales trends.
4. **Machine Learning Models**:
   * **Linear Regression**: For predicting sales based on variables such as store location, product category, or price.
   * **Random Forest or Gradient Boosting**: Use these ensemble methods to predict liquor sales based on various factors such as store characteristics, time of year, promotions, and weather data.
   * **XGBoost**: A powerful gradient boosting algorithm that can be used to predict sales based on multiple input features.
5. **Classification**:
   * **Sales Category Prediction**: If you're interested in predicting whether sales will be high, low, or medium, you can treat this as a classification problem. Use features like store location, product category, and past sales to classify future sales into predefined categories.
   * **Clustering**: Use clustering techniques (e.g., K-means) to group stores based on sales patterns, identifying underperforming stores or areas with untapped potential.
6. **Evaluation**:
   * **Regression Metrics**: Use Mean Absolute Error (MAE), Mean Squared Error (MSE), and R-squared to evaluate regression models predicting sales quantities.
   * **Time Series Metrics**: Evaluate your time-series models using the mean absolute percentage error (MAPE) or RMSE (Root Mean Squared Error).

**Step 4: Ethical Considerations:** What are some potential ethical concerns of this topic or analyzing the data?

1. **Privacy**:
   * The dataset does not contain personally identifiable information, but it’s important to ensure compliance with privacy laws when using any supplementary data that might involve individuals.
2. **Bias**:
   * Ensure that the model does not inadvertently favor certain stores or cities over others based on the dataset's distribution. This could happen if the sales data from one region is much larger or more frequent than from other areas.
3. **Fair Representation**:
   * The dataset should represent various types of products and stores. Avoid creating models that exclusively cater to specific product categories (e.g., spirits) or large stores.

**Step 5: Challenges/Issues**- What are some issues and challenges do you think you might face?

1. **Data Sparsity**:
   * If certain store locations or product categories have fewer sales data points, you might struggle with sparse data. You may need to aggregate data or focus on higher-selling stores/products.
2. **Seasonality**:
   * Liquor sales are highly seasonal, with peaks during holidays or warmer months. This makes predicting sales challenging as you must account for such patterns in your model.
3. **Real-Time Predictions**:
   * If you're building a real-time recommendation or prediction system, ensuring your model can handle dynamic data and provide real-time predictions will be a challenge.

**Step 6: References:**

* Iowa Liquor Control Board Dataset
* "Time Series Forecasting with Prophet" by Facebook.

**About this Dataset**

**Context**

The Iowa Department of Commerce requires that every store that sells alcohol in bottled form for off-the-premises consumption must hold a class "E" liquor license (an arrangement typical of most of the state alcohol regulatory bodies). All alcoholic sales made by stores registered thusly with the Iowa Department of Commerce are logged in the Commerce department system, which is in turn published as open data by the State of Iowa.

This dataset contains information on the name, kind, price, quantity, and location of sale of sales of individual containers or packages of containers of alcoholic beverages.

This dataset is relatively straightforward, but one source of further information on the contents of the data is [this Gist](https://gist.github.com/dannguyen/18ed71d3451d147af414).

**Inspiration**

This data is probably a representative sample of sale activity for alcohol in the United States, and can be used to answer many questions thereof, like: how much alcohol is sold and consumed in the United States? What kind? What are the most popular brands and labels? What are the most popular mixers? What is the distribution of prices paid in-store? Etcetera.