**MBAN 5110 Final**

**Runmeng Qin, Wei Lu, Yuhui Gu**

**Data Preparation**

We import the dataset provided: data\_set\_hackathon.csv, and check the information about that. And then we convert all currency value to EUR, convert ‘items’ to ‘int64’, convert 'order\_date’ and 'requested\_delivery\_date' to ‘datetime’ that convivence for the following analysis. Also calculate a new column for ‘Lead Time’. We find that there do not have any missing value and have some outliers (8) in 'Value EUR' and 'items'. We cut them and the final total row in dataset is 2412.

**EDA**

The distribution of order values is right-skewed, indicating that most order values are on the lower end, with fewer orders having high values. The peak of the distribution occurs around 50 EUR with approximately 175 counts. Similarly, the number of items ordered per transaction is right-skewed. The peak of this distribution appears to be around 8 items with 175 counts. The lead time distribution seems to be normally distributed, peaking around the mean, but it also shows some irregularities, such as negative lead times, which might indicate data entry errors or returns. The scatter plot suggests a positive correlation between the number of items ordered and the order value, which is to be expected.

**Feature Engineering**

We extract year and month from Order Date and Requested Delivery Date and change the different objects in 'Customer Country Code', 'Route', and 'Product Code’ to numeric that will be used in the following analysis of the finding in correlation.

**Time series**  
We use Time Series to examine whether seasonality influences the probabilities of choice.

The Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) plots have been generated for the first 12 lags, which can help us identify the appropriate parameters for the SARIMA model. Additionally, the Augmented Dickey-Fuller (ADF) test was performed to check the stationarity of the time series, with the following results:

* Test Statistic: Approximately -4.17
* p-value: Approximately 0.00073588

The test statistic being significantly negative and the p-value being close to 0 suggests that the time series is stationary and does not require differencing (d=0). This is an important step before fitting a SARIMA model, as it requires the series to be stationary.  
Next, we conduct a comparative analysis of two models, ARIMA and SARIMA, to determine which performs more effectively. Based on the AIC values, the simpler SARIMA model, which incorporates a seasonal component, appears to offer a superior fit. The model was trained using historical sales data, excluding the most recent 12 months, and is utilized to generate forecasts that extend beyond the range of the available historical data.

To analyze seasonality, we decompose the time series data into three distinct components: trend, seasonal, and residual. The trend component shows a gradual increase in the number of parka items ordered, followed by a decrease. This suggests that there was a period of growth followed by a decline over the observed time frame. The seasonal component shows clear spikes at regular intervals, indicating a strong seasonal pattern in parka sales. The residual component graph shows that for the most part, the residuals hover around zero, which is good, but there are a few significant outliers. These outliers could be due to unexpected events or anomalies not accounted for by the seasonal or trend components.

In summary, the decomposition indicates that parka sales have a strong seasonal pattern, as expected for a seasonal product like parkas.

**Multiple Linear Regression**

First, we use a correlation matrix of the dataset is calculated and visualized using a heatmap to understand the relationships between different variables. Key features like 'Value EUR', 'Order Month', 'Order Year', 'Customer Country Code Numeric', 'Product Code Numeric', and 'Route Numeric' are selected for predicting the 'Items' variable.

And then, we use Multiple Linear Regression model that was built using these features. The dataset is split into training and testing sets, with the model trained on the training set and predictions made on the test set. The model's performance is evaluated using the mean squared error (MSE) and R² score, which are 14.48 and 0.62 respectively. The intercept is 2482.7, which is the baseline 'Items' when all features are zero. Each coefficient shows the change in 'Items' per unit increase in a feature, keeping others constant. 'Value EUR' and 'Order Month' positively influence 'Items', increasing by 0.07 and 0.14 units respectively per unit rise. Conversely, 'Order Year' decreases 'Items' by about 1.23 units per year. 'Customer Country Code' and 'Product Code' slightly increase 'Items', while 'Route Numeric' reduces it, indicating diverse impacts of these factors on predicting 'Items'.

Next, we use the scatter plot to show and check the relationship between the actual and predicted number of items, with points clustered around the dashed line of perfect prediction, indicating a reasonably good model fit (the features we chose are correct and valuable for the following model’s requests), though some variance and outliers are present.

**Choice model:**

We chose the logistic regression model as the foundation for our choice model. The data was split into training and testing subsets to develop this model, which predicts the order month for each observation. A classification report was generated to assess model performance. Unfortunately, the model did not perform exceptionally well, achieving an accuracy score of 0.35. Among the 12 months, the model performed best in predicting December, with a correct labeling probability of 0.53. Additionally, this model calculates the probabilities of each possible outcome for each observation, providing insight into the likelihood of orders occurring in each month. We also produced a confusion matrix to offer an overview of the labeling accuracy.

For a more accessible view of how observations are distributed across each month, based on the calculated probabilities, we created several columns to display the probability associated with each month.

Furthermore, we employed a random forest regressor to forecast the individual monthly demand for the upcoming 12 months. This predictive model will assist the company in preparing for next year's supply.