Introduction

The T-100 Domestic Market (All Carriers) dataset is a comprehensive collection of domestic market data reported by both U.S. and foreign air carriers. Encompassing essential information such as carrier details, origin and destination airports, service classes, and passenger, freight, and mail statistics, this dataset is a valuable resource for analyzing air travel within the United States and its territories. The dataset spans from the year 1990 to 2023, providing a historical perspective on the dynamics of the air transportation industry.

Key Dataset Properties:

1. Records: The dataset comprises a vast collection of 7,224,651 records, capturing a wide array of air travel information.

2. Fields: With 36 fields, the dataset covers a comprehensive range of attributes, including details about carriers, airports, passenger counts, freight, and mail.

3. Frequency: The data is reported every month, allowing for a detailed examination of trends and variations over time.

Key Terms and Definitions:

1. Air Freight: Property, other than express and passenger baggage, transported by air.

2. Airline ID: An identification number assigned by the US Department of Transportation (DOT) to uniquely identify an airline or carrier. It remains constant even if the carrier changes its code, name, or holding company.

3. Airport Code and Airport ID: The three-character alphanumeric code issued by the U.S. Department of Transportation for official airport designation, with corresponding unique identification numbers for airports.

4. Carrier Code: A code assigned by the International Air Transport Association (IATA) to identify a carrier. This code may not always be unique as it can be assigned to different carriers over time.

5. City Market ID: An identification number assigned by the DOT to identify a city market, used for consolidating airports serving the same city market.

6. FIPS (Federal Information Processing Standards): Codes assigned to various geographic entities to simplify data collection, processing, and dissemination by the Federal Government.

7. Market (Using T100 Data): Passenger, freight, and/or mail data that enplane and deplane between two specific points with the same flight number, defining a market. A new market begins if the flight number changes.

8. Mile: A statute mile (5,280 feet) used as the unit for all mileage computations.

9. Passenger: Any person on board a flight who is not a member of the flight or cabin crew.

10. Unique Carrier, Unique Carrier Entity, and Unique Carrier Name: Identifiers to distinguish carriers and entities, including codes and names most recently used.

11. World Area Code (WAC): Numeric codes used to identify geopolitical areas like countries, states, provinces, territories, or possessions within the various data banks maintained by the Office of Airline Information (OAI).

This dataset provides a wealth of information for analyzing trends, patterns, and insights into the domestic air travel market, making it a valuable tool for researchers, analysts, and industry professionals.

Project Scope

For this project, we have focused our analysis on the T-100 Domestic Market (All Carriers) dataset, specifically covering the data from the year 2018 to August 2023. This timeframe allows us to concentrate on recent years, offering insights into the trends and dynamics of the domestic air travel market over this period.

We propose to tackle the business problem of market share prediction through supervised regression analysis. Accurate market share predictions are crucial for businesses and industries as they try to gain a competitive advantage (for example, resource optimization) and make data-driven decisions. By developing a predictive machine learning model, we aim to support business organizations with actionable insights, to help them sharpen their market presence and stay on top of their industries.

In a fast-paced business environment, staying competitive is a constant challenge. Accurate market share predictions can provide a strategic advance for corporates, by assisting them to identify growth opportunities and assessing the impact of other important factors. By focusing on this problem, we can maximize machine learning and data advantages to help inform organizational decisions in the current marketplace.

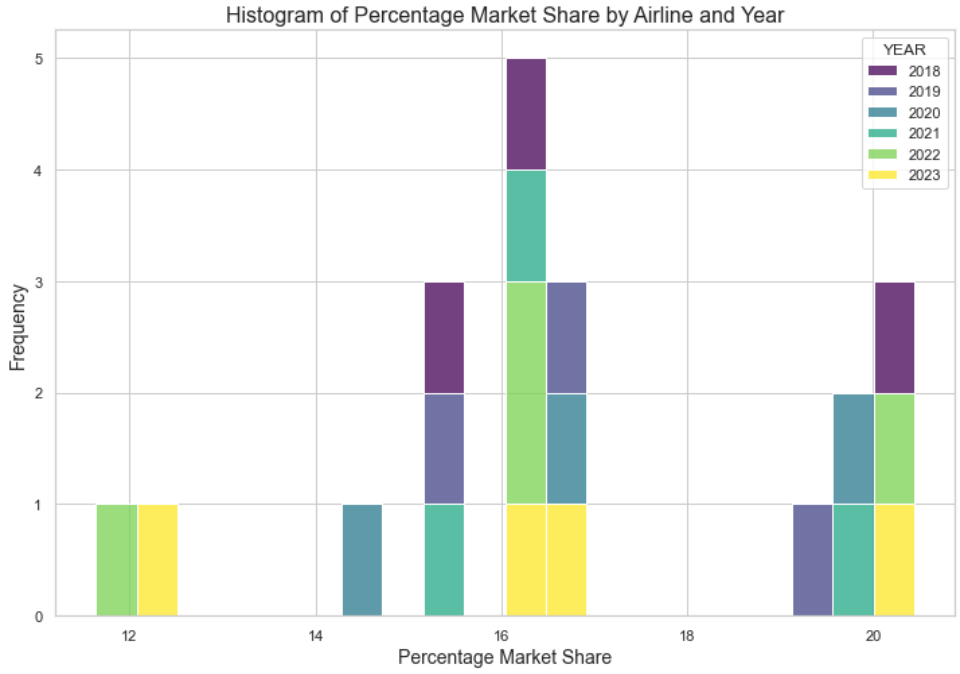
Data Cleaning and Exploration

The dataset was filtered to include relevant classes (A, C, E, F, G) for analysis. Which gives us the idea for scheduled flights only.
Cleaning involved removing rows with missing values, ensuring the dataset's integrity.
Market share for each airline was calculated based on the total passengers, offering insights into the market dominance of individual carriers based on “AIRLINE_ID”.

In [8]:
```
for i, df in enumerate(dfs, start=1):
    new_set = df[df['CLASS'].isin(['A', 'C', 'E', 'F', 'G'])]

    new_set_cleaned = new_set.dropna(axis=0)

    total_passenger = new_set_cleaned.groupby('AIRLINE_ID')['PASSENGERS'].sum()

    # market share for each carrier
    market_share = total_passenger / total_passenger.sum()

    df = df.merge(market_share.rename('MARKET_SHARE'), left_on='AIRLINE_ID', right_index=True)

    df['LOG_MARKET_SHARE'] = np.log(df['MARKET_SHARE'] + 1)

    globals()[f'df{i}'] = df
    globals()[f'new_set{i}'] = new_set_cleaned
```

Iterative loops were used to calculate “MARKET SHARE LAG” for each data frame (df2 to df6) based on the previous year's market share (df1). The 'MARKET_SHARE_LAG' column was created, representing the change in market share from the previous year.

Looking at the top 10 “Market Share Percentage”, one can see how it changes.

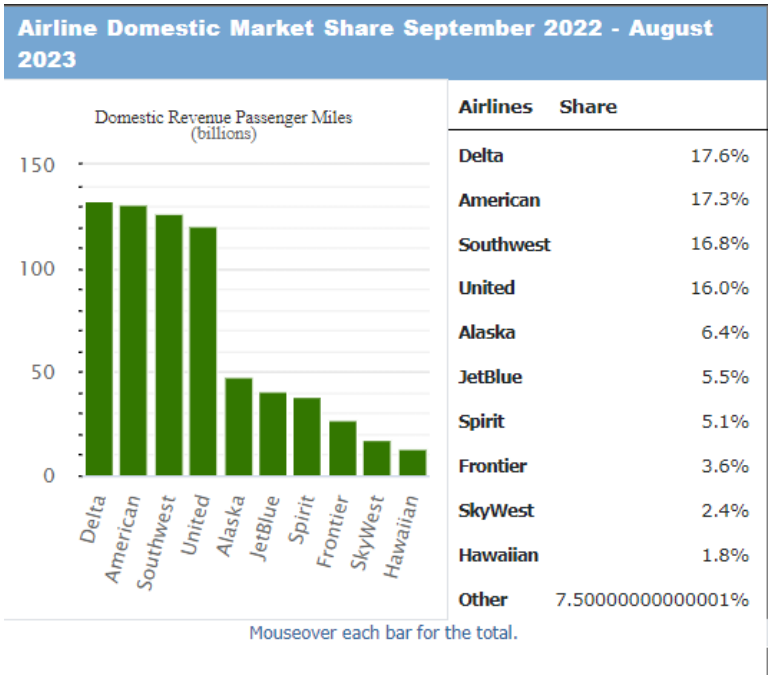


The top 20 Market Share percentages for different carriers can be seen below.

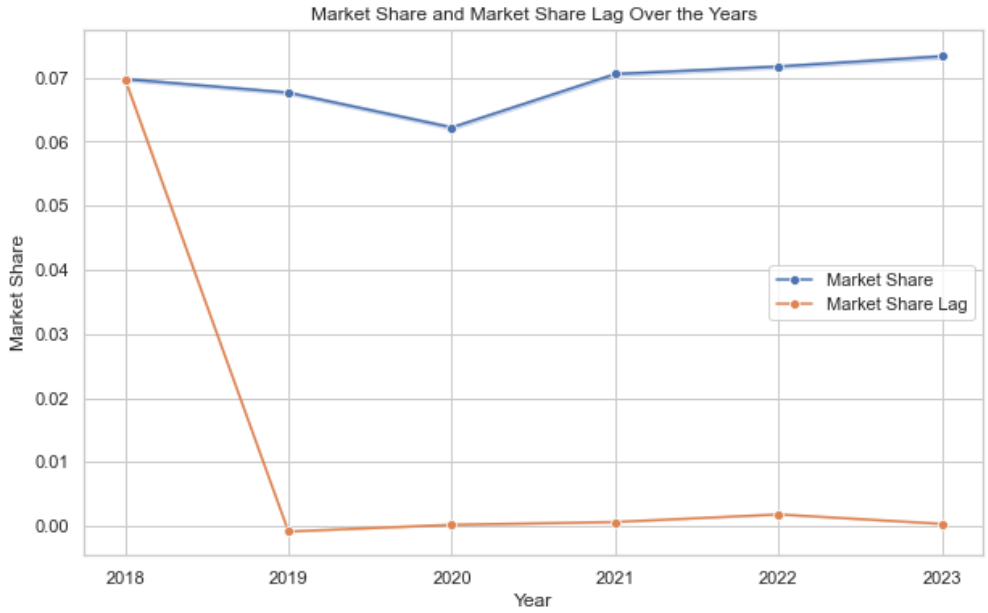
	AIRLINE_ID	UNIQUE_CARRIER_NAME	PERCENTAGE_MARKET_SHARE	YEAR
0	19393.0	Southwest Airlines Co.	20.443644	2018
4	19393.0	Southwest Airlines Co.	20.402985	2022
5	19393.0	Southwest Airlines Co.	20.218701	2023
3	19393.0	Southwest Airlines Co.	19.919350	2021
2	19393.0	Southwest Airlines Co.	19.844465	2020
1	19393.0	Southwest Airlines Co.	19.520593	2019
66	19805.0	American Airlines Inc.	16.885632	2020
59	19790.0	Delta Air Lines Inc.	16.784464	2019
63	19790.0	Delta Air Lines Inc.	16.754150	2023
67	19805.0	American Airlines Inc.	16.372126	2021
62	19790.0	Delta Air Lines Inc.	16.283758	2022
58	19790.0	Delta Air Lines Inc.	16.256818	2018
69	19805.0	American Airlines Inc.	16.104567	2023
68	19805.0	American Airlines Inc.	16.056294	2022
65	19805.0	American Airlines Inc.	15.529737	2019
64	19805.0	American Airlines Inc.	15.356788	2018
61	19790.0	Delta Air Lines Inc.	15.248618	2021
60	19790.0	Delta Air Lines Inc.	14.379180	2020
96	19977.0	United Air Lines Inc.	12.505096	2023
95	19977.0	United Air Lines Inc.	11.636331	2022

For our calculation, we have considered the number of Passengers as the main criteria to determine carrier/airline Market Share.

Below is the comparison of Market Share according to the official website, which considers the Market Share according to the Revenue that is being achieved by the Carrier.

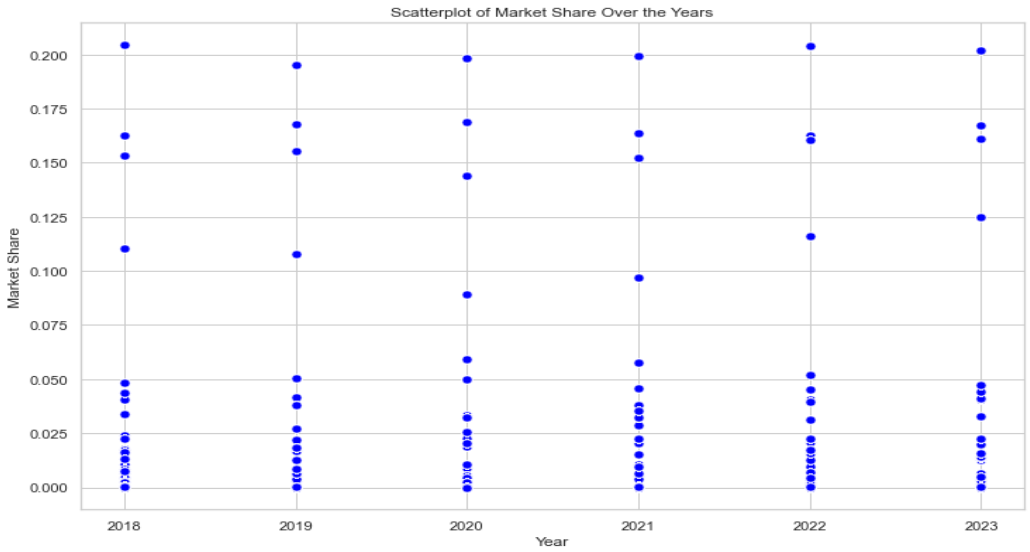


Looking at the calculation one can see a similarity in the results, which gives us the idea of the direction of this project, of using the data for the prediction of Passenger Airline Market Share. In a general economic theory, the revenue highly depends on the customer in this case passenger. The 4 top Airlines that come are Southwest, Delta, American, and United. The same that we can see in the revenue as well.

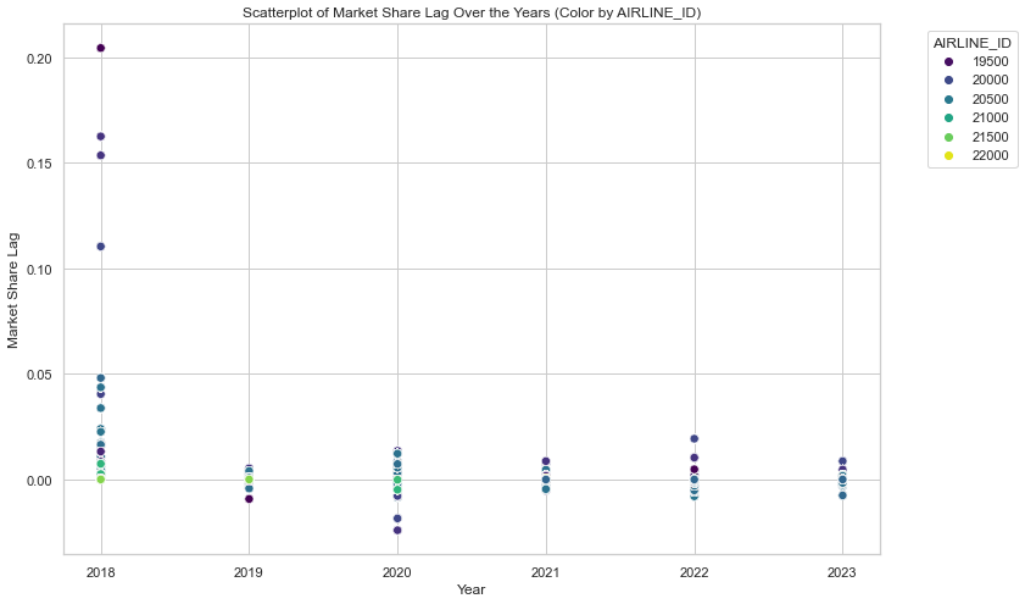


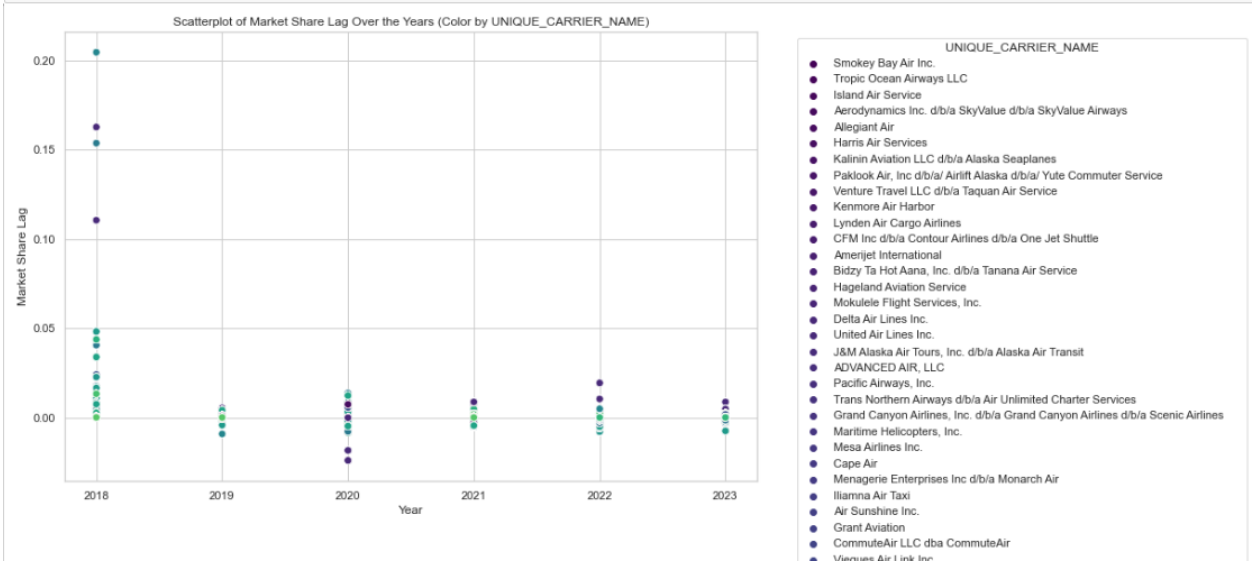
The above gives us an understanding of how the market share changes, from 2018 to 2023. Now Market Share lag represents the difference between the two consecutive market share differences for airlines and by taking the average of those market share we can see how it changes. For this analysis we considered data from 2018 hence that's the benchmark of the Market share that we are considering.

The below scatterplot shows how the Market Share has been distributed over the years.



The below two graphs give us an idea of how the Market Share lag behaves (ignoring 2018), the difference seems to be evident for 2020, which was the year of the COVID pandemic and Airlines were highly affected by that which can be seen in the above graphs as well.

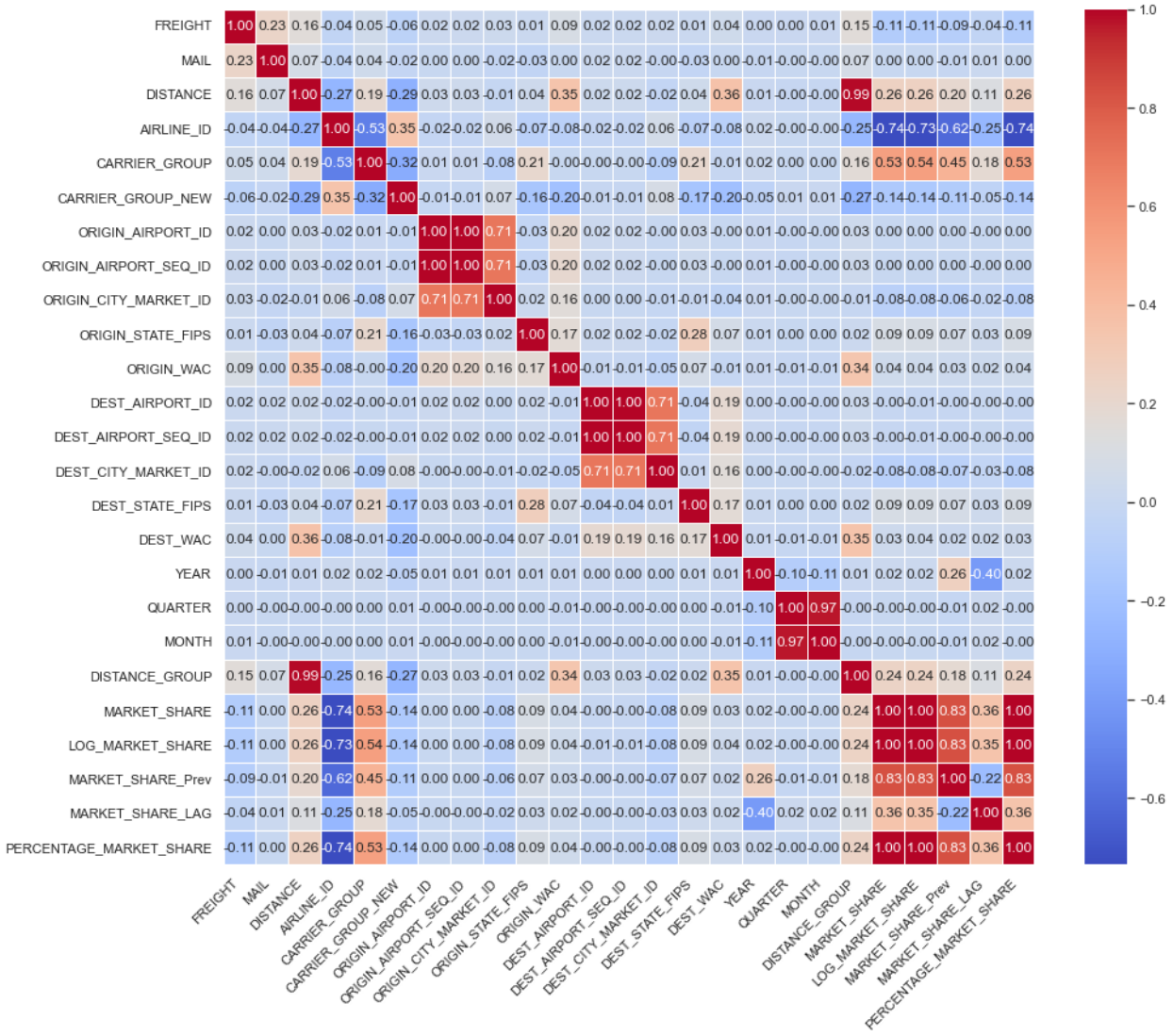




Correlation Analysis

A correlation matrix was generated to explore the relationships between variables, focusing on identifying features strongly correlated with log market share. Now basic understanding of Market and feature selection is also required. Just by looking at the number if we consider the

highly positive and negative values then we might end up taking the wrong feature for further model training.



When we calculate Market Share we have already used passengers, hence we do not include that in the feature selection. Now, let us understand how market share can be predicted using other features to understand if a newer service can be added or if there are more flights required that can be helpful in the growth, hence we haven't included the highly correlated AIRLINE ID.

The features that are being considered are, ['CARRIER_GROUP', 'FREIGHT', 'DISTANCE', 'ORIGIN_WAC', and 'DEST_WAC']. Brief information about the columns:

'CARRIER_GROUP' is the international code for the carriers, for example, 0 is for International carriers, 1 is for Regional Carriers (including Large, Medium, Commuter, and Small Certified), and many more.

FREIGHT: Property, other than express and passenger baggage, transported by air in pounds. So how many of these are being carried out from one place to another?
DISTANCE: From start to end how much distance that flight takes. This is important as the Freight and other revenue depends on this.
ORIGIN_WAC and DEST_WAC: Flight origin and destination World codes. This variable gives us an idea of how an airline might work in between those two destinations.

Absolute correlation values can be seen below

```
MARKET_SHARE               1.000000
PERCENTAGE_MARKET_SHARE    1.000000
LOG_MARKET_SHARE           0.999805
MARKET_SHARE_Prev          0.831359
AIRLINE_ID                 0.735650
CARRIER_GROUP              0.534361
MARKET_SHARE_LAG           0.355436
DISTANCE                   0.255944
DISTANCE_GROUP             0.239540
CARRIER_GROUP_NEW          0.137180
FREIGHT                    0.111972
DEST_STATE_FIPS            0.087661
ORIGIN_STATE_FIPS          0.085796
DEST_CITY_MARKET_ID        0.082503
ORIGIN_CITY_MARKET_ID      0.075882
ORIGIN_WAC                 0.038243
DEST_WAC                   0.034033
YEAR                       0.021291
DEST_AIRPORT_ID            0.004980
DEST_AIRPORT_SEQ_ID        0.004979
QUARTER                    0.002996
MONTH                      0.002224
ORIGIN_AIRPORT_SEQ_ID      0.001138
ORIGIN_AIRPORT_ID          0.001138
MAIL                       0.001113
Name: MARKET_SHARE, dtype: float64
```

Now looking at the absolute value we can see how different variables are correlated with Market Share. Generally speaking, highly positive and negative values are considered highly correlated, but it does not signify that the variables might be used for prediction. Many variables are correlated to one another for example Distance, Distance_group both provide similar information, and keeping one makes more sense to avoid the variable bias.

Prediction Models

Using different regression models, we tried to predict the Market Share for different routes based on the above-mentioned features. To start with the prediction the criteria that we chose was to divide data from 2018 to 2021 to be used as a training data set and for testing data set we considered 2022 and 2023 data. Having historical data gives us a better understanding of the model.

Simple Linear Regression

Now once we have the training and testing sets. We further split into features and target variables for training and testing.

The result we get is the R square 33.5% with a Mean Square Error to be 0.0014. This shows that approximately 34% of the variability in market share can be explained by the selected features in your model and a lower MSE indicates better model performance.

```
r_squared = r2_score(y_test, predictions)

mse = mean_squared_error(y_test, predictions)

print(f'R-squared (R²): {r_squared}')
print(f'Mean Squared Error (MSE): {mse}')
```

```
R-squared (R²): 0.3352543747034419
Mean Squared Error (MSE): 0.004155309115112647
```

The performance seems to be good but can be improved by using other techniques.

Lasso and Ridge Regression

Ridge and Lasso regression are regularization techniques that aim to improve the performance of linear regression models by preventing overfitting and handling multicollinearity.

Lasso Regression Results:
R-squared: 0.073 (approximately 7.3% of variability explained).
Coefficients: [0, -1.60e-08, 2.03e-05, 0, 0].
Intercept: 0.0502.

Ridge Regression Results:
R-squared: 0.335 (same as linear regression).
Coefficients: [0.0462, -1.71e-08, 1.51e-05, -1.81e-05, -2.59e-05].
Intercept: -0.0561.

Linear Regression and Ridge Regression provide similar R-squared values, suggesting that Ridge does not significantly improve the model in this case.
Lasso Regression, however, yields a lower R-squared, indicating a reduction in explanatory power. This is expected as Lasso introduces sparsity by driving some coefficients to zero.

```
0.07307924457855108

(array([ 0.00000000e+00, -1.59946871e-08,  2.03181829e-05, -0.00000000e+00,
        -0.00000000e+00]),
 0.05020538390466017)
```

```
print(lasso_grid.score(X_test, y_test))
```

```
0.3347214174405684
```

```
ridge_grid.fit(X_train, y_train)
```

```
0.3352543618592816
```

Gradient Boosting Regression

Gradient Boosting is an ensemble learning technique that combines the predictions of multiple weak learners (typically decision trees) to create a stronger predictive model. It builds trees sequentially, with each new tree correcting the errors of the previous ones.

We get the R-squared as 0.5481 (approximately 54.81% of variability explained) and the MSE comes out to be 0.00282

```
print(f"Gradient Boosting Regression R-squared: {gb_reg.scor
```

```
Gradient Boosting Regression MSE: 0.002824770276801512
Gradient Boosting Regression R-squared: 0.5481073412463447
```

Gradient Boosting Regression outperforms both Linear Regression and Lasso Regression in terms of R-squared and MSE. The higher R-squared indicates that the gradient-boosting model captures more variability in the target variable. The lower MSE suggests that the predictions are closer to the actual values, resulting in a more accurate model.

Random Forest Regression

The Random Forest Regression model was trained using 100 decision trees, and the random_state parameter was set to 42 for reproducibility.

The MSE is relatively low at 0.00158, indicating that the model's predictions are close to the actual values on average and a high R-squared value of 0.747 suggests that the Random Forest Regression model captures a significant portion of the variability in the market share data.

```
Random Forest Regression MSE: 0.001582123350333686
Random Forest Regression R-squared: 0.7468997981428536
```