The Role of Network Design in Airline Performance

A Data-Driven Approach

Padma Sree Manda

Department of ITDS, University of North Texas

DSCI 5900: Special Problems

Dr. Sourav Chatterjee

12/01/2024

I would like to thank Harley Moulder for his guidance throughout the research and writing of this paper. Those insightful discussions significantly enhanced my understanding about the airlines industry and the quality of my work.

The Role of Network Design in Airline Performance

A Data-Driven Approach

Abstract

This study explores the influence of airline network structures on operational efficiency and on-time performance, with a focus on Hub-and-Spoke and Point-to-Point systems. Utilizing data from the Bureau of Transportation Statistics, the analysis assesses key performance indicators, including delays, load factors, and stage lengths. The findings indicate that while Hub-and-Spoke systems perform well in terms of connectivity and scalability, they are susceptible to delay propagation. In contrast, Point-to-Point systems emphasize passenger convenience but encounter higher costs per route.

To explore these dynamics, the study employs regression analysis, decision tree models, and network analysis techniques. Network centrality measures emphasize the significant roles of major hubs in legacy carrier operations, while highlighting the more distributed approach typical of low-cost carriers. The findings yield insights into the trade-offs between efficiency, passenger convenience, and network resilience, ultimately informing strategies for optimizing airline performance.

Table of Contents

List of Tables

List of Figures

I. Introduction

One of the widely researched topics in the airline industry is route structure or different networks that can reduce the impact of delays. This paper aims to understand the underlying relationship between Network structure and airline on-time performance.

The airline industry is characterized by its dynamic and competitive landscape, where network design is integral to operational efficiency and passenger satisfaction. Leading carriers in the United States—namely American Airlines, Delta Air Lines, United Airlines, and Southwest Airlines—demonstrate significant revenue generation through distinctive strategic approaches that are responsive to the evolving demands of travelers. These strategies are shaped by technological advancements, changing travel patterns, and varying economic conditions.

Legacy carriers, such as American, Delta, and United, predominantly utilize the **Hub-and-Spoke** network model. This architecture consolidates flights into central hub locations, thus facilitating connectivity across expansive networks with a reduced number of routes. Mainline aircraft, generally equipped with over 100 seats, are instrumental within this structure, optimizing passenger capacity while promoting efficient travel routes linking major metropolitan areas (IBISWorld, 2024). The operational scale afforded by this model enables legacy airlines to capitalize on economies of scale, thereby offering a comprehensive array of international and domestic services. Nevertheless, these carriers face mounting challenges due to escalating fuel costs, inflation, and transformative travel trends, notably the transition from business to leisure travel spurred by remote work paradigms (IBISWorld, 2024).[2]

In contrast, Southwest Airlines exemplifies the efficacy of the **Point-to-Point** network model, which establishes direct connections between cities, eliminating the reliance on centralized

[2] Accessed through the University of North Texas Library's IBISWorld subscription service. (last accessed on 27th November 2024)

hubs. This model prioritizes passenger convenience and reduced travel times while leveraging low-cost, no-frills fare structures (IBISWorld, 2024). In a bid to maintain competitiveness, legacy carriers have increasingly adopted basic economy fare configurations, thereby emulating the budget-friendly appeal traditionally associated with low-cost carriers. Additionally, major airlines are expanding their loyalty programs and encouraging consumer loyalty through various mechanisms such as credit card partnerships, mileage benefits, and subscription packages (IBISWorld, 2024).

Regional airlines serve as a vital complement to these network models, catering to smaller airports and less densely populated regions, often functioning as subsidiaries of larger carriers. These airlines play a crucial role in enhancing connectivity in underserved markets and providing entry-level employment opportunities within the aviation sector (IBISWorld, 2024). The industry's resurgence following the pandemic in 2024 is bolstered by an improving consumer sentiment, which in turn stimulates demand for both regional and mainline services.

This study aims to investigate the implications of the Hub-and-Spoke and Point-to-Point models on operational performance and passenger convenience. Employing datasets from the Bureau of Transportation Statistics, the analysis critically evaluates key metrics, including delays, load factors, and stage lengths, across both legacy and low-cost carriers. Utilizing advanced methodologies, such as regression analysis and network modeling, the study seeks to illuminate the advantages and constraints inherent to each network design.

By examining the operational trade-offs associated with these models, this research aims to enrich the understanding of how airlines can optimize their network configurations, mitigate costs, and enhance passenger satisfaction amid the evolving dynamics of the market landscape.

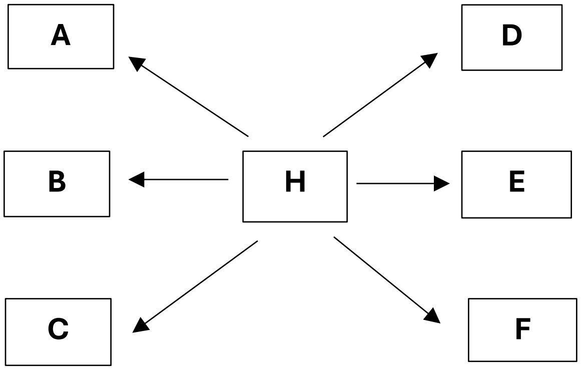
II. **Literature Review**

The paper investigates and compares two primary types of airline network structures: the "Hub-and-Spoke" system and the "Point-to-Point" system. The analysis is framed within the context of airline operational efficiency, passenger convenience, and overall network performance.

A. ***Hub-and-Spoke***

In the Hub-and-Spoke model, airlines operate flights from numerous outlying airports (spokes) into a central hub. Passengers from different origins converge at the hub and then continue to their final destination, often requiring a transfer. This configuration is widely used by legacy carriers due to its operational and economic advantages.

Figure 1: Hub – and – Spoke Network Structure



Benefits -

- **Operational Efficiency:** This model minimizes the total number of routes required to connect multiple cities, reducing operational complexity and capital investment in aircraft fleets (Button, 2002) (Gillen & Morrison, 2005).
- **Market Power:** Dominance at hub airports often enables airlines to achieve higher market shares and pricing power, attracting passengers with comprehensive connectivity options (Borenstein, 1989).

- **Economies of Scale:** High traffic density through hubs allows airlines to use larger aircraft, reducing per-passenger costs and increasing profitability.
- **Increased Network Scope:** Adding new cities to the network involves connecting them to the hub, which can potentially provide access to all existing destinations in the network.

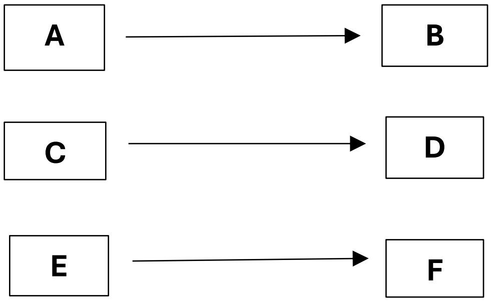
Drawbacks -

- **Delay Propagation:** Adverse weather or operational disruptions at the hub can cascade delays across the network, significantly impacting passengers and operations (Franke, 2004).
- **Longer Travel Times:** Layovers and transfers increase total travel time compared to direct flights, reducing passenger convenience.
- **Resource Intensity:** Hub operations require substantial investments in infrastructure, staff, and facilities to manage connecting passengers and peak traffic periods (Donoghue, 2002).

B. Point-to-Point

The Point-to-Point model operates direct flights between different cities without a central hub. Passengers can travel directly to their destinations without the need for transfers. Low-cost carriers (LCCs), such as Southwest Airlines, often employ this model to provide flexible, direct service.

Figure 2: Point – to – Point Network Structure



Benefits –

- **Passenger Convenience:** By eliminating layovers, the model offers shorter travel times, aligning with passenger preferences for direct flights (Cook & Goodwin, 2008).
- **Flexibility:** Airlines can dynamically adapt routes based on demand, allowing for cost-efficient route planning and scheduling.
- **Reduced Hub Dependence:** Airlines mitigate the risks associated with maintaining and relying on large-scale hub operations.

Drawbacks -

- **Limited Connectivity:** This model often struggles to efficiently connect low-demand city pairs, limiting its application to high-traffic routes.
- **Higher Costs per Route:** The lack of consolidated traffic flows leads to lower load factors, potentially increasing operational costs per passenger.

The literature emphasizes that although the “Hub-and-Spoke” system provides extensive connectivity and operational scalability, it is inherently complex and susceptible to disruptions. In contrast, the “Point-to-Point” model is more streamlined and passenger-friendly, yet it sacrifices network breadth and operational economies of scale (Nasrollahi & Kordani, 2023). Cook and Goodwin (2008) point out that no single system is ideal; most airlines implement hybrid models to strike a balance between efficiency and flexibility.

Recent studies emphasize incorporating passenger preferences, such as minimizing delays and maximizing convenience, into network design. Multiobjective optimization models suggest that balancing airline costs with passenger satisfaction is pivotal for long-term operational success

(Nasrollahi & Kordani, 2023). Innovations in fleet planning, dynamic scheduling, and the integration of advanced analytics further shape network strategies in modern aviation.

III. Data Description

A. Reporting Carrier On-Time Performance

The data is collected from the "Bureau of Transportation Statistics" which has related data from the year 1987 till the present date (Bureau of Transportation Statistics, n.d.).[3] For analysis purposes, data has been restricted to 2018 yearly data. The grain of the data is per carrier per city pair per flight per date. Each row contains daily information about flights operated by different carriers across various city pairs in the United States. For this project, we will concentrate on three prominent legacy carriers: American Airlines, Delta Airlines, and United Airlines, along with Southwest Airlines, recognized as a low-cost carrier. Below are some key details provided by the data.

1. Date information: - Variables like Day, Month and Year.
2. Carrier information: - Unique Carrier code for each airline
3. Flight information: - Flight number, Origin, and Destination related variables.
4. Performance metrics: – Various variables such as Dep_delay, Arr_delay, Carrier_delay, Cancelled, Diverted, etc., are available to analyze each carrier's performance.
5. Other metrics – Variables such as Dep_time_blk, Arr_time_blk, Taxi_out, Taxi_in, Wheel_off, Wheels_on, distance, etc.

[3] For more information, see the Bureau of Transportation Statistics website at https://www.transtats.bts.gov/DL_SelectFields.aspx?gnoyr_VQ=FGJ&QO_fu146_anzr=b0-gvzr (last accessed November 18, 2024).

In addition to the variables present in the dataset, several key performance indicators have been established to facilitate data analysis. The Departure Delay Ratio assesses the proportion of a flight's departure delay relative to the scheduled elapsed time (CRS_ELAPSED_TIME). A higher ratio indicates that the departure delay constitutes a significant portion of the overall planned flight duration.

Flight Time Efficiency evaluates how effectively the flight covered the scheduled distance. Ideally, this ratio should be close to 1, signifying that the actual flight time closely aligns with the scheduled time (CRS_ELAPSED_TIME). A lower value suggests inefficiency, indicating that the flight took longer in the air than anticipated.

Congestion at Origin gauges the level of crowding at an airport during the flight's departure. This can be quantified by counting the number of flights scheduled to depart from the same airport within the same time block (e.g., during the same hour).

Legacy carriers typically operate under a hub-and-spoke network structure, while Southwest Airlines employs a point-to-point network model. To examine the impact of these differing network structures on on-time performance, a dummy variable was created. In this analysis, Arrival Delay (in minutes) serves as the target or dependent variable, with various performance metrics used as predictors or independent variables.

B. T-100 Segment (Traffic Data)

The data is collected from the "Bureau of Transportation Statistics" which has related data from the year 1990 till the present date (Bureau of Transportation Statistics, n.d.).[4] For analysis purposes, data has been restricted from 2018 to 2023 yearly data. This data is about the traffic of

[4] For more information, see the Bureau of Transportation Statistics website at https://www.transtats.bts.gov/DL_SelectFields.aspx?gnoyr_VQ=GDM&QO_fu146_anzr=Nv4%20Pn44vr45 (last accessed November 18, 2024).

various carriers over the years. The grain of the data is per carrier per city pair per month per year. Each row contains detailed information regarding flight traffic data, including payload and seat availability, for each carrier across various city pairs in the United States. The dataset employs four different carriers, similar to the reporting carrier's on-time performance data. It offers insights into the following aspects:

1. Date Information – fields like month and year
2. Carrier Information – Carrier code and carrier name
3. Flight Information – variables such as flight number, origin and destination.
4. Traffic data – Payload, Seats, Freight, Mail, Distance, Ramp_Time, Air_Time.
5. Other Fields – Distance Group, Departures_schedules, Departures_Performed.

In addition to the aforementioned variables, various performance indicators were derived from the existing data metrics. One such indicator, Available Seat Miles (ASM), serves as a quantifiable measure of an airline's overall capacity. It reflects the total number of seat miles available for passengers, taking into account both the number of seats offered and the distance flown. This metric is crucial for evaluating operational efficiency and capacity utilization within the airline industry.

$$\text{ASM} = \text{Seats} \times \text{Distance (Miles)}$$

Revenue Passenger Miles (RPM) serves as a critical metric that quantifies the actual traffic carried by an airline, specifically reflecting the distance traveled by paying passengers. This measure provides valuable insights into the operational performance and market demand for air travel services offered by the airline.

$$\text{RPM}=\text{Passengers} \times \text{Distance(Miles)}$$

The Load Factor is a quantitative metric that assesses the proportion of available seats occupied by paying passengers on an airline. This metric is pivotal for evaluating the capacity utilization of an airline, as it reflects the efficiency and effectiveness of an airline's operations in maximizing revenue from its available seating inventory. A higher load factor indicates a more efficient use of capacity, while a lower load factor may suggest underutilization, impacting overall profitability.

$$\text{Load Factor}=(\text{RPM}/\text{ASM})\times 100$$

Stage length refers to the mean distance traversed during each individual flight within an airline's operational network. This metric serves as an important indicator of the spatial extent of the airline’s services, providing insights into the characteristics and dynamics of its flight routes.

$$\text{Stage Length}=\text{Distance}/\text{Departures\_Performed}$$

Block Time Efficiency serves as a metric for evaluating the effectiveness of time allocation during flight operations, specifically contrasting the duration spent airborne against the time allocated for ground movements, including taxiing in and out. A higher Block Time Efficiency value signifies a greater proportion of time spent in the air relative to the comprehensive duration of the flight.

$$\text{Block Time Efficiency}=\text{Air\_Time}/\text{Ramp\_to\_Ramp}$$

IV. Models

In this project, we developed several models, including Linear Regression and Decision Tree Regression, to analyze and predict the target variable, Arrival Delay, based on a range of predictors. The dataset comprises approximately 3.8 million rows and 64 features, of which 54 predictors were utilized in the model-building process by partitioning the data into training and testing sets.

A. Feature Selection Process

The selection of features has been conducted through a combination of statistical methodologies necessitating human oversight and an automated machine-learning model. This dual approach ensures a robust and comprehensive framework for analysis.

Statistical Linear Regression is a quantitative modeling technique that employs the Ordinary Least Squares (OLS) method to minimize the sum of squared residuals, thus identifying the optimal regression line (Statsmodels Developers, n.d.).[5] The implementation of this technique is facilitated through the "Statsmodels" package in Python, which allows for the fitting of the model to the training dataset and generates a comprehensive summary report. This report delineates crucial metrics such as the R-squared value, Adjusted R-squared value, coefficients for each predictor, and their respective p-values.

In this analysis, four distinct regression models have been constructed, each utilizing different sets of predictors. These models have yielded valuable insights into the significance of each predictor in relation to the overall model performance. In accordance with statistical best practices, the presence of collinearity among the predictors has been assessed, leading to the exclusion of one predictor to enhance model validity. Subsequent to the evaluation of the summary report, predictors identified as statistically significant—characterized by a p-value of less than 0.05—have been selected as features for the development of Machine Learning models. This rigorous selection process ensures that the predictive framework is both robust and relevant.

[5] For more information, see the Statsmodels documentation at https://www.statsmodels.org/dev/generated/statsmodels.regression.linear_model.OLS.html (last accessed November 18, 2024).

SequentialFeatureSelector from the “mlxtrend” package is utilized alongside Linear Regression for predictor selection (Raschka, n.d.).[6] In this approach, forward selection is employed, using the R-squared value as the criterion for evaluation. Consequently, the model explores various combinations of predictors to identify the set that yields the highest accuracy for the linear regression model.

B. Linear Regression

LinearRegression model is used from the “sklearn.linear_model” package of Python to develop a regression model that predicts Arr_delay based on selected features (Scikit-learn Developers, n.d.).[7] The objective of this machine-learning model is to identify the optimal set of model parameters (coefficients) that minimizes the prediction error on the training dataset. This process is frequently accomplished through the application of optimization algorithms, and the problem is commonly referred to as the “Least Squares” optimization problem. The objective function in this context delineates the error or cost associated with the model's predictive performance. In the realm of linear regression, the Mean Squared Error (MSE) serves as the predominant objective function. MSE quantifies the mean of the squared differences between the predicted values and the actual target values within the training dataset, thereby providing a measure of the model’s accuracy.

Two Linear Regression models were constructed utilizing distinct sets of features, one comprising 50 predictors and the other consisting of 12 predictors. The analysis using the OLS method, including R-squared values and the correlation matrix, indicated a clear linear relationship

[6] For more information, see the MLxtend documentation at https://rasbt.github.io/mlxtend/api_subpackages/mlxtend.feature_selection/#sequentialfeatureselector (last accessed November 18, 2024).

[7] For more information, see the Scikit-learn documentation at https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html (last accessed November 18, 2024).

between the target variable and the predictors. As a result, the outcomes of the models appear promising and will be elaborated upon in the Results Section.

C. Decision Tree Regressor

DecisionTreeRegressor model is used from the "sklearn.tree" package of Python to develop a regression model that predicts Arr_delay based on selected features (Scikit-learn Developers, n.d.).[8] Unlike Linear Regression, Decision Tree Regressor is a model that can be used irrespective of the presence or absence of a linear relationship between the target and predictor variables.

The Decision Tree is a prominent machine learning model that functions through the recursive partitioning of input data into distinct subsets based on the values of its intrinsic features. The algorithm identifies the features that most effectively segregate the data into separate groups at each stage, resulting in a tree-like structure delineating the decision rules. A key metric employed in decision tree algorithms is entropy, which serves to measure impurity or disorder within the data subsets throughout the process of constructing the tree. Entropy quantifies the level of uncertainty or randomness intrinsic to a set of data points and is integral to the computation of information gain (Li, 2019).[9]

Entropy reaches its apex when the dataset is evenly distributed across the various classes. Information Gain itself is a metric that quantifies the reduction in uncertainty concerning the target variable (or class labels) achieved by partitioning the data based on a specific attribute (Li, 2019).[7] The underlying premise is to select the attribute that results in the most substantial reduction of uncertainty, or entropy, within the resulting subsets. Attributes exhibiting higher information gain

[8] For more information, see the Scikit-learn documentation at https://scikit-learn.org/dev/modules/generated/sklearn.tree.DecisionTreeRegressor.html (last accessed November 18, 2024).

[9] For more information, see the Towards Data Science article at https://towardsdatascience.com/https-medium-com-lorrli-classification-and-regression-analysis-with-decision-trees-c43cdbc58054 (last accessed November 18, 2024).

are prioritized for node splitting within decision trees due to their capacity to provide more significant insights and to effectively diminish overall uncertainty in the dataset.

Nevertheless, decision tree models are susceptible to the phenomenon of overfitting, as they may produce excessively intricate trees that adhere closely to the training data, even incorporating noise and fluctuations inherent within it. Such complexity often yields trees with numerous branches and nodes, each tailored to particular training samples, thereby achieving commendable performance on the training dataset. However, these overly complex trees typically fail to generalize effectively, potentially leading to inadequate performance when applied to novel, unseen data.

To mitigate the risk of overfitting, various techniques are employed in the construction of decision trees. These include constraining the maximum depth of the tree, pruning the tree structure, and establishing minimum sample sizes for splits. By simplifying the tree structure and promoting improved generalization, these methodologies enhance the decision tree's capacity to deliver accurate predictions on previously unencountered data (Baladram, 2024).[10] In this regard, the parameter max_depth serves as an early stopping criterion, imposing limitations on the depth or complexity of the tree during the training phase. In essence, max_depth stipulates the maximum number of levels from the root to the leaf nodes, and judiciously setting an appropriate value can effectively prevent excessive depth and the associated risk of overfitting.

The model has been optimized by tuning various values of max_depth to identify the most effective combination of training and testing scores. Presented below are charts illustrating the relationship between model complexity and both training and testing scores across multiple feature

[10] For more information, see the Towards Data Science article at https://towardsdatascience.com/decision-tree-regressor-explained-a-visual-guide-with-code-examples-fbd2836c3bef (last accessed November 18, 2024).

sets. These visualizations provide insight into the model's generalization capabilities and performance metrics.

Figure 3: Model Complexity Chart for Predictors based on Sequential Feature Selector

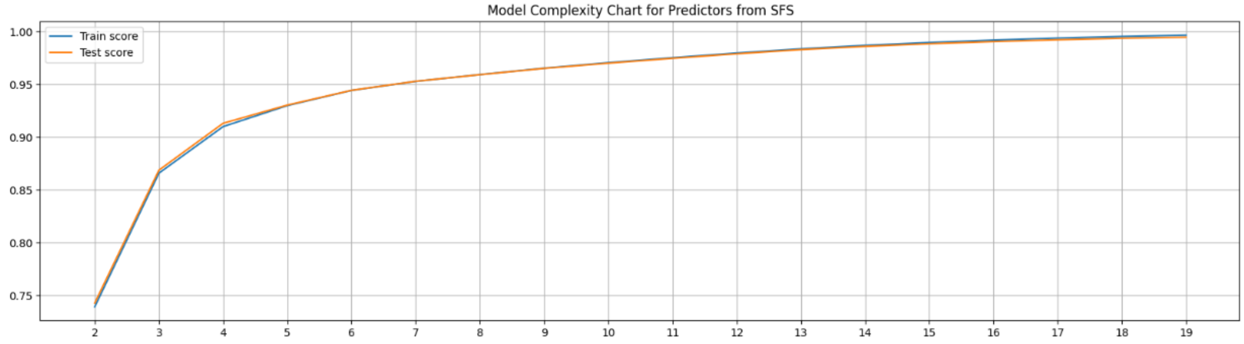
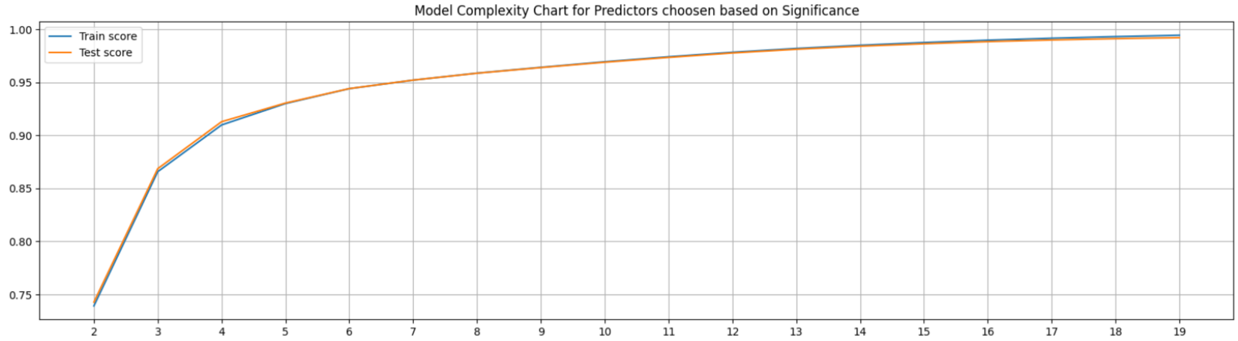


Figure 4: Model Complexity Chart for Predictors based on Statistical Significance



According to the charts, it is evident that the scores of the decision tree regressor remain quite consistent, regardless of the feature selection method employed. Additionally, it is worth noting that there is little variation between the training and testing scores in both scenarios. Therefore, a max_depth value of 13 has been chosen, as it yields high scores and indicates a good fit. The evaluation of the model indicates that the score for the value of 13 is sufficiently high to yield a robust R-squared value. Furthermore, the observed differences between the training and testing scores suggest that the model is effectively generalizing to unseen data.

D. Network Analysis

A network analysis has been conducted on the dataset utilizing the “networkx” package to uncover the underlying patterns within the data for each airline. For each airline, a weighted directed network graph was created, with airports represented as nodes and the connections between them as edges. The number of trips between city pairs serves as the weight for these edges. To gain insights into the structure of each network graph, several measures have been employed, which are detailed below.

Degree Centrality is a measure that quantifies the potential connections or edges a node may have within a network graph (NetworkX Developers, n.d.).[11] In essence, a higher degree centrality indicates that an airport has more direct connections to other airports. Consequently, hub airports exhibit a greater degree centrality value among legacy carriers, which is consistent with the findings.

Betweenness Centrality is a measure of how frequently a node appears on the shortest path between all pairs of nodes within a network (NetworkX Developers, n.d.).[12] This implies that in a network of airline carriers, hubs typically have a higher value, as they are crucial components of the network with as many direct connections to other airports as possible.

Closeness Centrality measures the average shortest distance from a node to all other nodes in a network, reflecting how swiftly it can connect with the rest of the network (NetworkX

[11] For more information, see the NetworkX documentation at https://networkx.org/documentation/stable/reference/algorithms/generated/networkx.algorithms.centrality.degree_centrality.html (last accessed November 18, 2024)

[12] For more information, see the NetworkX documentation at https://networkx.org/documentation/stable/reference/algorithms/generated/networkx.algorithms.centrality.betweenness_centrality.html (last accessed November 18, 2024).

Developers, n.d.).[13] In essence, nodes that contribute to efficient pathways within the network will exhibit higher values of this measure.

Eigenvector centrality is a measure of the influence that a node wields within a network. In essence, nodes connected to other significant nodes tend to have a higher value (NetworkX Developers, n.d.).[14] This metric provides a more nuanced understanding of a node's importance by taking into account not only its connections but also the quality of those connections. In the context of network structure, it helps identify airports that hold substantial influence due to their connectivity to other major hubs, thereby shaping the overall flow of air traffic and connectivity patterns.

[13] For more information, see the NetworkX documentation at https://networkx.org/documentation/stable/reference/algorithms/generated/networkx.algorithms.centrality.closeness_centrality.html (last accessed November 18, 2024).

[14] For more information, see the NetworkX documentation at https://networkx.org/documentation/stable/reference/algorithms/generated/networkx.algorithms.centrality.eigenvector_centrality.html (last accessed November 18, 2024).

Table 1: Ten Important Nodes in the American Airlines Network and Their Measures

American Airlines	**Degree Centrality**	**Betweenness Centrality**	**Closeness Centrality**	**Eigen Vector Centrality**
DFW	1.722772	0.513241	0.878261	0.358076
CLT	1.287129	0.265092	0.737226	0.317693
ORD	1.000000	0.066508	0.668874	0.301547
PHX	0.950495	0.088152	0.655844	0.270662
PHL	0.772277	0.053351	0.619632	0.252991
MIA	0.712871	0.029666	0.608434	0.250079
LAX	0.633663	0.026989	0.594118	0.224689
DCA	0.316832	0.001446	0.543011	0.161102
JFK	0.316832	0.003759	0.540107	0.152014
MCO	0.198020	0.000114	0.526042	0.129164

Table 2: Ten Important Nodes in the Delta Airlines Network and Their Measures

DELTA	Degree Centrality	Betweenness Centrality	Closeness Centrality	Eigen Vector Centrality
ATL	1.805556	0.757290	0.911392	0.393214
MSP	0.951389	0.125392	0.657534	0.325111
SLC	0.722222	0.077594	0.610169	0.276329
DTW	0.687500	0.035743	0.602510	0.274285
LAX	0.458333	0.025607	0.564706	0.219813
JFK	0.444444	0.011440	0.562500	0.216381
SEA	0.437500	0.042134	0.562500	0.208434
BOS	0.250000	0.001110	0.533333	0.167687
LGA	0.208333	0.001415	0.516129	0.114545
CVG	0.208333	0.000728	0.527473	0.137492

Table 3: Ten Important Nodes in the United Airlines Network and Their Measures

UNITED	Degree Centrality	Betweenness Centrality	Closeness Centrality	Eigen Vector Centrality
DEN	1.336449	0.284525	0.748252	0.337344
ORD	1.327103	0.264428	0.748252	0.343656
IAH	1.177570	0.222370	0.694805	0.312418
EWR	0.990654	0.149832	0.664596	0.291120
SFO	0.943925	0.111625	0.648485	0.285073
IAD	0.654206	0.036561	0.591160	0.235433
LAX	0.373832	0.026774	0.545918	0.168774
LAS	0.299065	0.001959	0.529703	0.159267
CLE	0.168224	0.000300	0.511962	0.111360
PHX	0.149533	0.000000	0.509524	0.116355

Table 4: Ten Important Nodes in the Southwest Airlines Network and Their Measures

Southwest	Degree Centrality	Betweenness Centrality	Closeness Centrality	Eigen Vector Centrality
DEN	1.428571	0.095731	0.777778	0.235965
DAL	1.404762	0.097442	0.770642	0.233489
MDW	1.404762	0.084622	0.770642	0.234011
LAS	1.309524	0.087827	0.743363	0.219166
BWI	1.285714	0.106798	0.736842	0.213824
HOU	1.214286	0.090879	0.717949	0.211213
PHX	1.214286	0.054254	0.717949	0.217286
MCO	1.142857	0.057511	0.700000	0.199511
STL	1.047619	0.036137	0.677419	0.203275
BNA	0.904762	0.027083	0.646154	0.184316

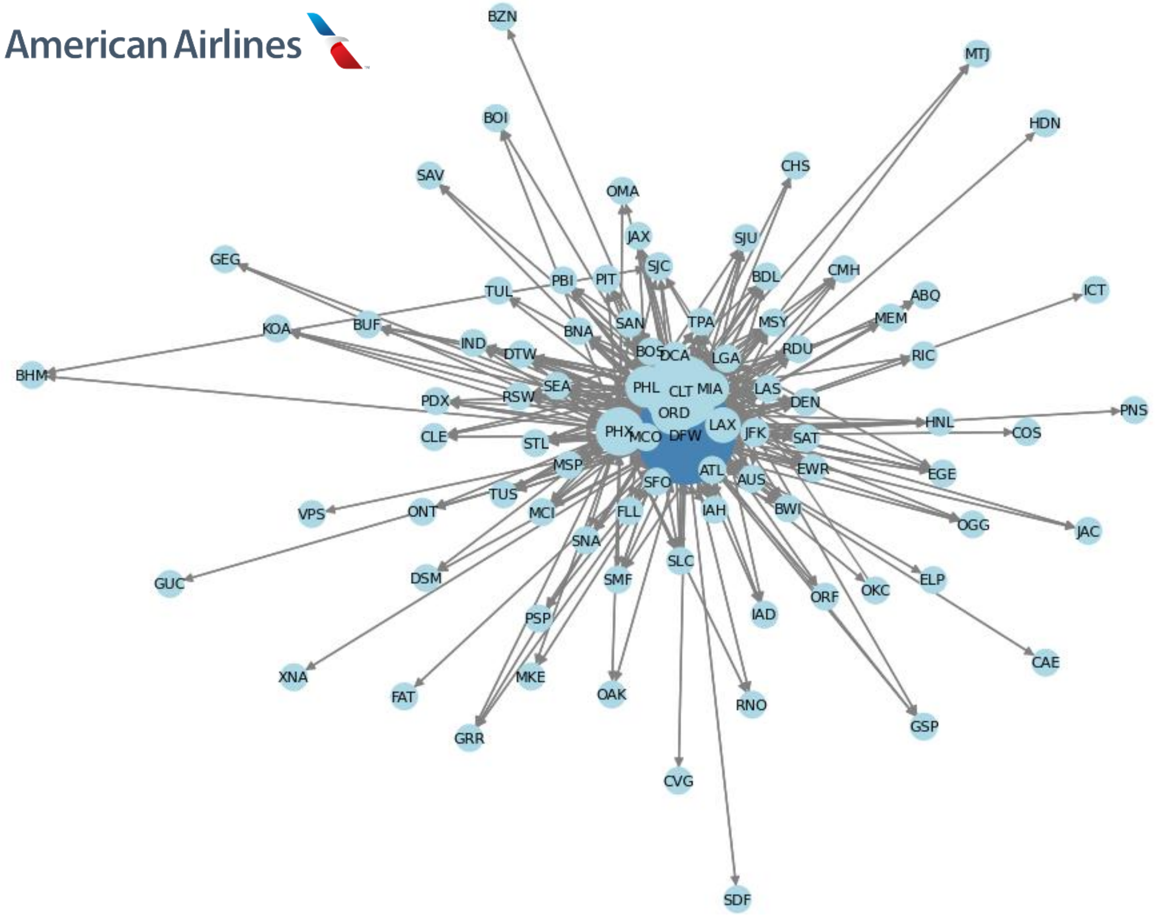
Legacy carriers are known to operate using a hub-and-spoke network structure. To illustrate this concept, ego-centric network graphs have been created for each legacy airline, with their hubs serving as the focal points (NetworkX Developers, n.d.).[15] The hubs represent the nodes with the highest degree centrality value. Additionally, betweenness centrality has been employed to indicate node size; thus, a higher betweenness centrality value corresponds to a larger node size.

The following analysis presents an ego-centric network of American Airlines, with Dallas/Fort Worth (DFW) serving as the ego and major hub. It is evident that this network operates

[15] For more information, see the NetworkX documentation at https://networkx.org/documentation/stable/reference/generated/networkx.generators.ego.ego_graph.html (last accessed November 18, 2024).

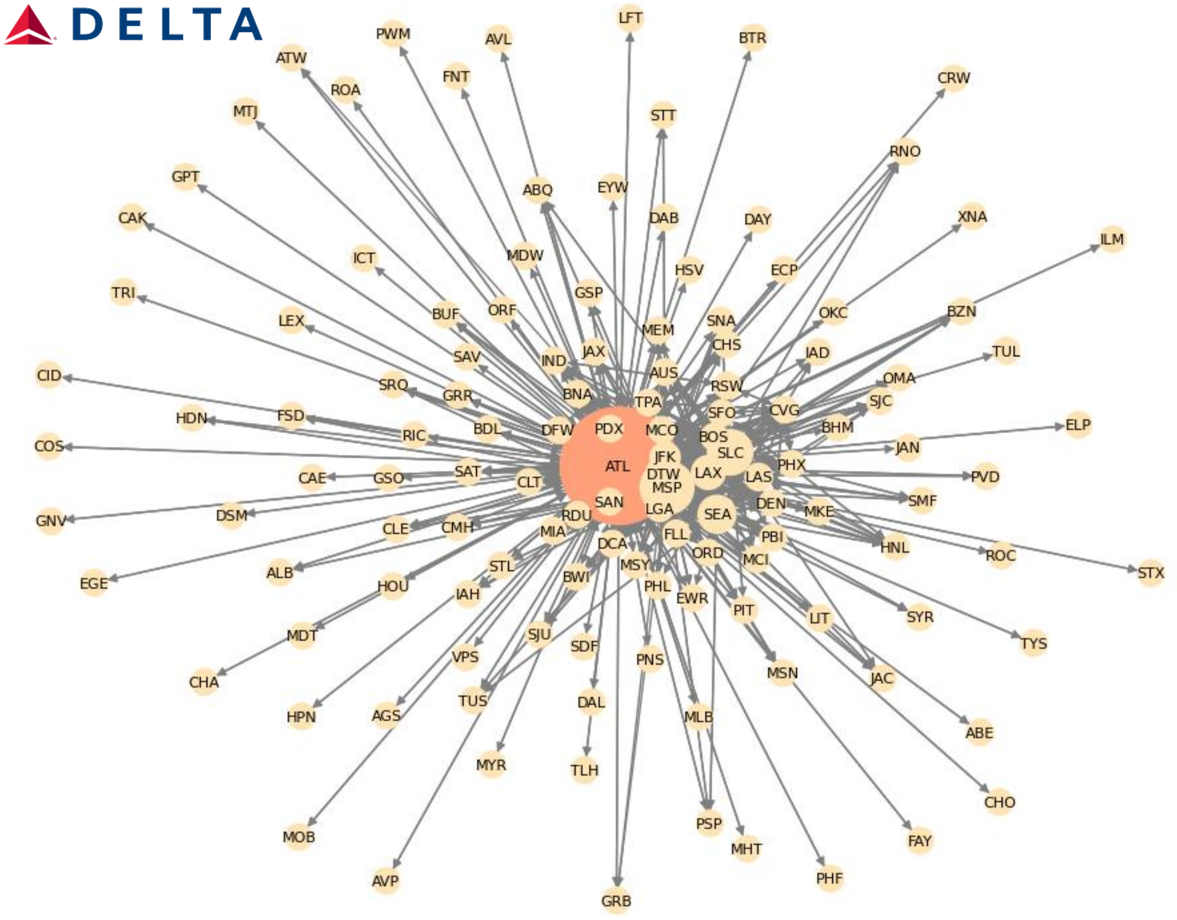
under a hub-and-spoke structure, featuring notable sub-hubs or important airports such as Philadelphia (PHL), Charlotte (CLT), and Chicago O'Hare (ORD). These sub-hubs are represented by larger node sizes, reflecting their higher betweenness centrality values compared to the spokes.

Figure 5: Ego-Centric Network Graph of American Airlines



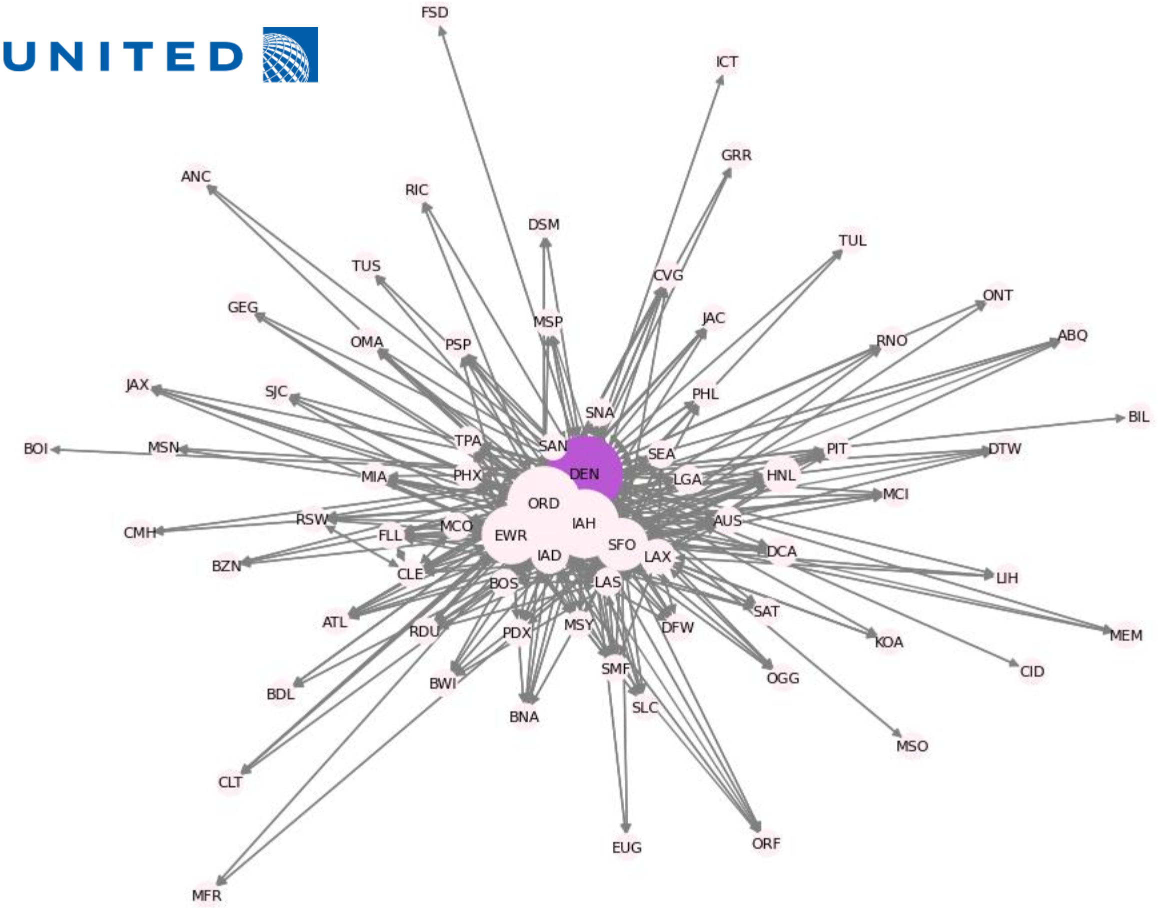
The following is an ego-centric network graph of Delta Airlines, with Atlanta (ATL) identified as the central hub. The structure of the network is clearly a hub-and-spoke model, where ATL serves as the primary hub and other airports function as the spokes. In contrast to American Airlines, only the hub displays a significantly larger node size, reflecting a higher betweenness centrality value. This indicates that Delta Airlines operates within a traditional hub-and-spoke network, anchored by a major hub while connecting to various other nodes.

Figure 6: Ego-Centric Network Graph of Delta Airlines



The following is an ego-centric network graph of United Airlines, with Denver (DEN) serving as the primary hub within the network. While it operates on a hub-and-spoke model, it differs from Delta in that it does not rely on a single hub. This is evident from the sizes of various airport nodes, which indicate that multiple airports—such as Chicago O'Hare (ORD), Newark (EWR), and Houston (IAH)—are also significant and play crucial roles in the network.

Figure 7: Ego-Centric Network Graph of United Airlines

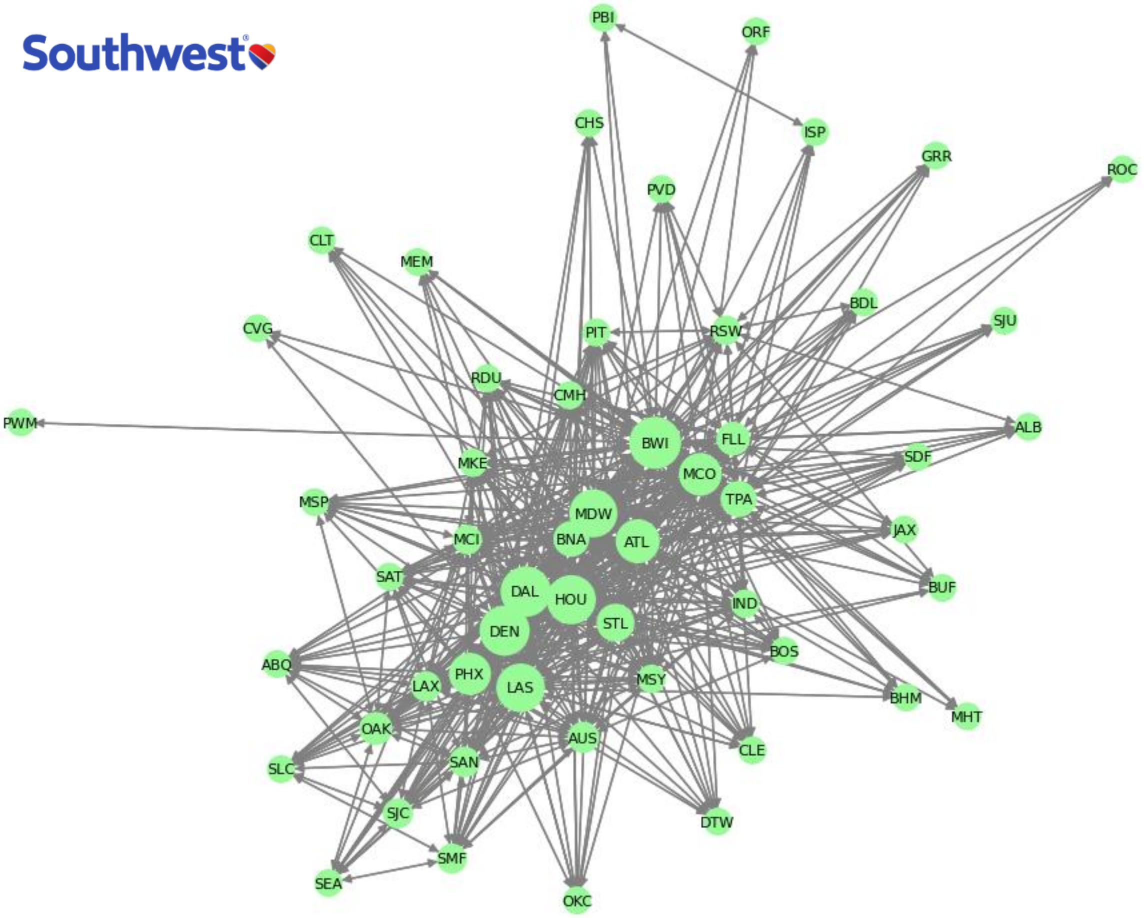


The following is an analysis of an ego-centric network graph pertaining to Southwest Airlines, utilizing Baltimore/Washington International (BWI) as the ego node. BWI is characterized by a higher degree centrality value, indicating its significance within the network. This graph illustrates that Southwest Airlines operates not under the traditional hub-and-spoke model, common among legacy carriers, but rather employs a point-to-point network strategy.

Furthermore, the relatively uniform sizes of the nodes, indicative of their degree of interconnectivity, suggest that airports such as Dallas Love Field (DAL), Baltimore/Washington International Airport (BWI), Denver International Airport (DEN), Chicago Midway International Airport (MDW), and Orlando International Airport (MCO) serve as high-traffic routes for Southwest Airlines.

This distinct operational model highlights the fundamental differences between Southwest Airlines and traditional legacy carriers, emphasizing a decentralized network structure that fosters greater flexibility and accessibility across its route system. The implications of this network design may also reflect on operational efficiency and customer service paradigms, further differentiating Southwest in the competitive airline industry.

Figure 8: Ego-Centric Network Graph of Southwest Airlines



V. Results

A. Statsmodels Results

Below are the results for the multiple linear regression which shows the summary report with values such as R square, Adjusted R square, AIC, BIC, Coefficient values, and their p-values. Hypothesis testing is involved in the report for the coefficient values.

H0 = Slope is zero

Ha = Slope is not equal to zero

We reject the null hypothesis if the p-value is less than 0.05, which means, the variables having a p-value less than 0.05 are statistically significant. Consequently, the variables having a p-value greater than 0.05 have been excluded from the model. The analysis has been refined to focus exclusively on the variables that exhibit statistical significance.

Table 5: Statistical Multiple Linear Regression Report

Model:	**OLS**	**Adj. R-squared:**	0.957
Dependent Variable:	ARR_DELAY	**AIC:**	27416383.7069
Date:	2024-11-11 17:55	**BIC:**	27416554.8004
No. Observations:	3840217	**Log-Likelihood:**	-1.3708e+07
Df Model:	12	**F-statistic:**	7.191e+06
Df Residuals:	3840204	**Prob (F-statistic):**	0.00
R-squared:	0.957	**Scale:**	73.803

	Coef.	**Std.Err.**	**t**	**P>|t|**	**[0.025**	**0.975]**
const	6.1787	0.0081	763.8021	0.0000	6.1628	6.1945
DEP_DELAY	17.9455	0.0179	1000.7501	0.0000	17.9104	17.9806
TAXI_OUT	4.2903	0.0053	815.2516	0.0000	4.2799	4.3006

TAXI_IN	2.4495	0.0047	515.7834	0.0000	2.4402	2.4588
DISTANCE	-3.0337	0.0054	-560.3664	0.0000	-3.0443	-3.0231
CARRIER_DELAY	13.1659	0.0118	1116.2769	0.0000	13.1428	13.1890
WEATHER_DELAY	6.1108	0.0067	915.8323	0.0000	6.0977	6.1239
NAS_DELAY	8.7844	0.0071	1234.0080	0.0000	8.7704	8.7983
SECURITY_DELAY	0.8075	0.0045	179.5076	0.0000	0.7987	0.8163
LATE_AIRCRAFT_ DELAY	12.3606	0.0113	1093.0995	0.0000	12.3385	12.3828
flight_time_efficiency	2.6378	0.0053	494.7016	0.0000	2.6274	2.6483
congestion_origin	0.2826	0.0046	62.0799	0.0000	0.2736	0.2915
NETWORK	-3.8344	0.0105	-365.3230	0.0000	-3.8550	-3.8138

Omnibus:	1448008.700	**Durbin-Watson:**	1.296
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	206514851.178
Skew:	-0.764	**Prob(JB):**	0.000
Kurtosis:	38.893	**Condition No.:**	8

B. Machine Learning Models

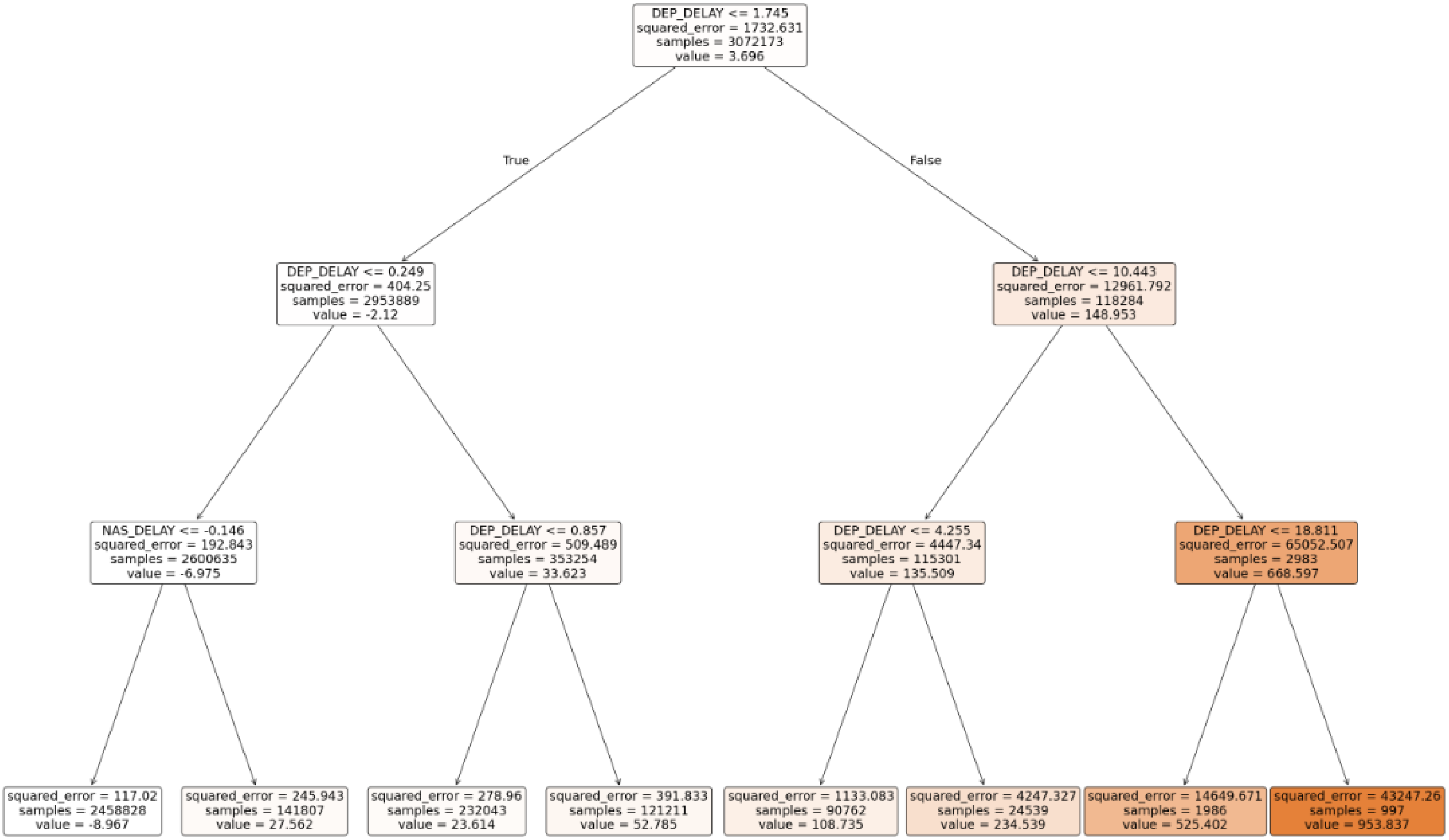
“Sequential Feature Selector” presents a ranked list of features that optimize the R-squared value through a forward selection methodology which is listed below in the table.

Table 6: List of Best Features in Forward Selection

DEP_DELAY	DISTANCE	CRS_ELAPSED_TIME	TAXI_IN
SECURITY_DELAY	NAS_DELAY	ACTUAL_ELAPSED_TIME	TAXI_OUT
WEATHER_DELAY	CARRIER_DELAY	LATE_AIRCRAFT_DELAY	NETWORK
departure_delay_ratio	congestion_origin	flight_time_efficiency	Dep_0001-0559
Dep_0600-0659	Dep_0700-0759	Dep_0800-0859	Dep_0900-0959
Dep_1000-1059	Dep_1100-1159	Dep_1300-1359	Dep_1400-1459
Dep_1500-1559	Dep_1600-1659	Dep_1700-1759	Dep_1800-1859
Dep_1900-1959	Dep_2000-2059	Dep_2100-2159	Dep_2200-2259
Arr_0001-0559	Arr_0600-0659	Arr_0700-0759	Arr_0800-0859
Arr_0900-0959	Arr_1000-1059	Arr_1100-1159	Arr_1200-1259
Arr_1300-1359	Arr_1500-1559	Arr_1600-1659	Arr_1700-1759
Arr_1800-1859	Arr_1900-1959	Arr_2000-2059	Arr_2100-2159
Arr_2200-2259	Arr_2300-2359		

The following is a tree plot representing the “decision tree model”, which showcases only the first three levels of depth. As detailed in the Models section, a decision tree operates by calculating entropy and the corresponding information gain. The feature that exhibits the highest information gain will serve as the root node, indicating its significance in predicting the Arrival Delay variable. From the tree plot, it is evident that Departure Delay is the root node and a key feature for prediction. This is followed by various types of delays, such as NAS Delay and Weather Delay, which appear further down the tree as child nodes.

Figure 9: Tree Plot for the Decision Tree Model



Below is a table depicting the various models developed and their corresponding train and test scores.

Table 7: R-square values of Different Models for Different sets of Features

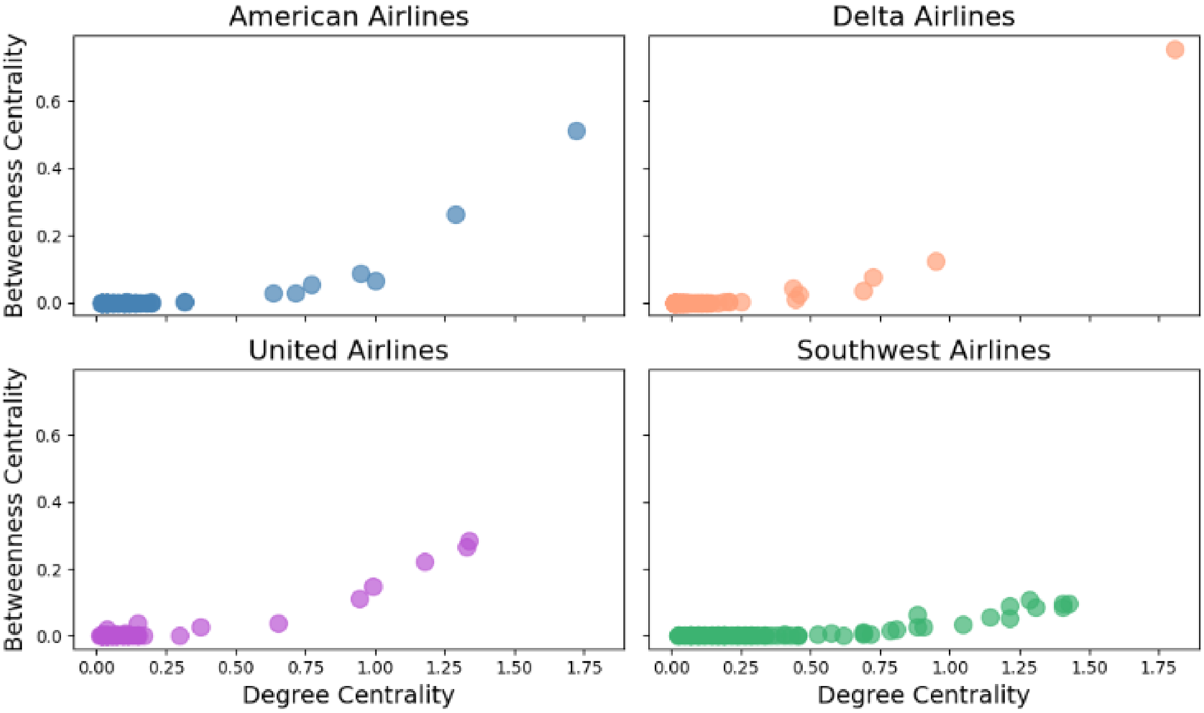
Model	Feature Selection Process	Train Score	Test Score
Linear Regression	Statistical (Human)	0.9572	0.9577
	Machine (Automatic)	0.9614	0.9620
Decision Tree Regressor	Statistical (Human)	0.9821	0.9813
(Max Depth = 13)	Machine (Automatic)	0.9836	0.9829

There is little distinction between the scores of models with features selected based on statistical significance and those using Sequential Feature Selector (SFS). It is worth noting that all models demonstrate a good fit, as the test scores are close to the training scores. This indicates that the models are performing well on unseen data and generalizing effectively. Based on the scores, the Decision Tree Regressor outperforms the Linear Regression model; however, the difference between the two is not substantial.

C. Network Analysis

The following analysis presents four subplots of scatter plots, each corresponding to a distinct airline. In these representations, Degree Centrality is plotted on the x-axis, while Betweenness Centrality is depicted on the y-axis. This visualization aims to elucidate the relationships and variations in centrality measures among the airlines under consideration.

Figure 10: Comparison plot of Airlines on Betweenness Centrality



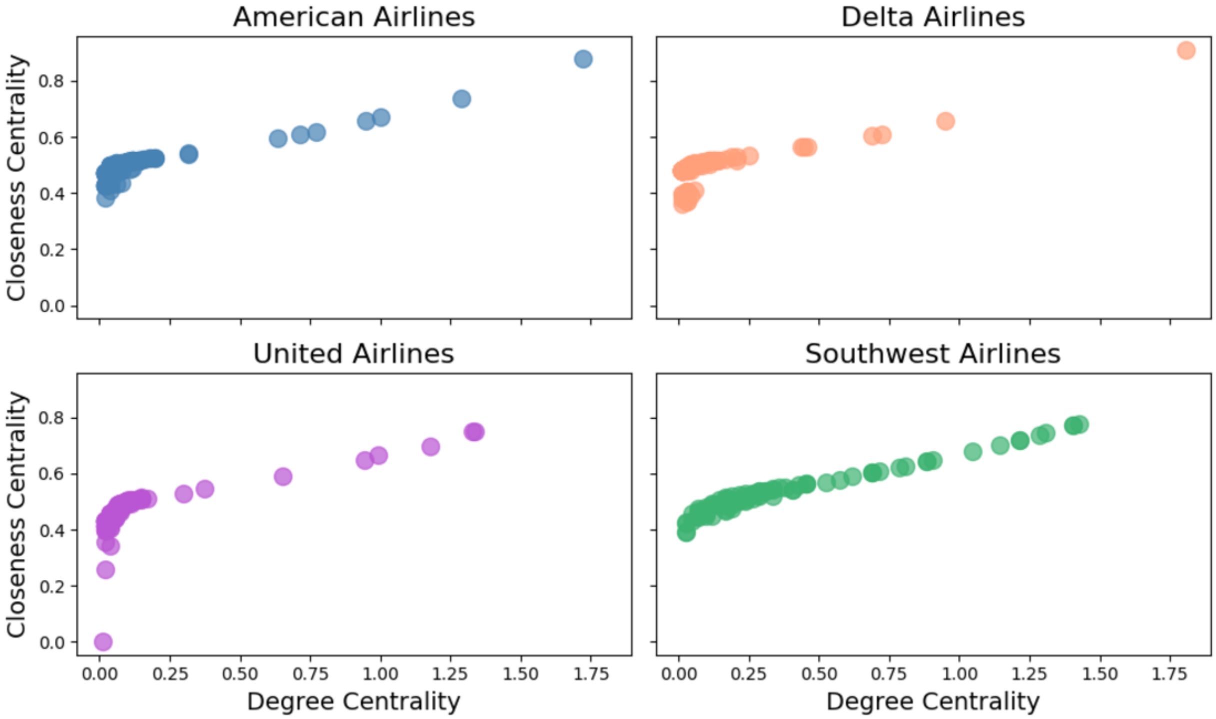
The first point to note is that Southwest Airlines has a significantly lower betweenness centrality value compared to legacy carriers. This indicates that there are fewer shortest paths traversing the airports served by Southwest. This difference may stem from the fact that high-traffic airports are interconnected rather than being organized around a central hub, which characterizes a hub-and-spoke system. In legacy carriers, airports with the highest degree centrality values also tend to have higher betweenness values, as these airports function as hubs for their respective airlines and maintain direct connections to various other airports.

Interestingly, Delta Airlines operates with a single major hub that exhibits the highest values for both degree and betweenness centrality when compared to other legacy carriers. In contrast, United Airlines has multiple hubs that have lower degree and betweenness centrality values in comparison to Delta's hub. This suggests that United Airlines relies on several major hubs, each connected to various other airports. American Airlines demonstrates a similar trend, featuring one major hub along with sub or regional hubs to enhance geographic connectivity. These nodes typically show mid-level degree centrality values but low betweenness centrality values.

A similar pattern is observed in Southwest Airlines, where the interconnected set of airports results in a high degree of centrality but not necessarily frequent inclusion in the shortest paths, resulting in a lower betweenness value. Nodes exhibiting both low betweenness and degree centrality are typically peripheral. For legacy carriers, these serve as spokes within the network structure, while for Southwest Airlines, they encompass airports with limited traffic and thus, low connectivity.

The following presentation consists of four subplots depicting scatter plots for each airline, with Degree Centrality represented along the x-axis and Closeness Centrality along the y-axis.

Figure 11: Comparison plot of Airlines on Closeness Centrality



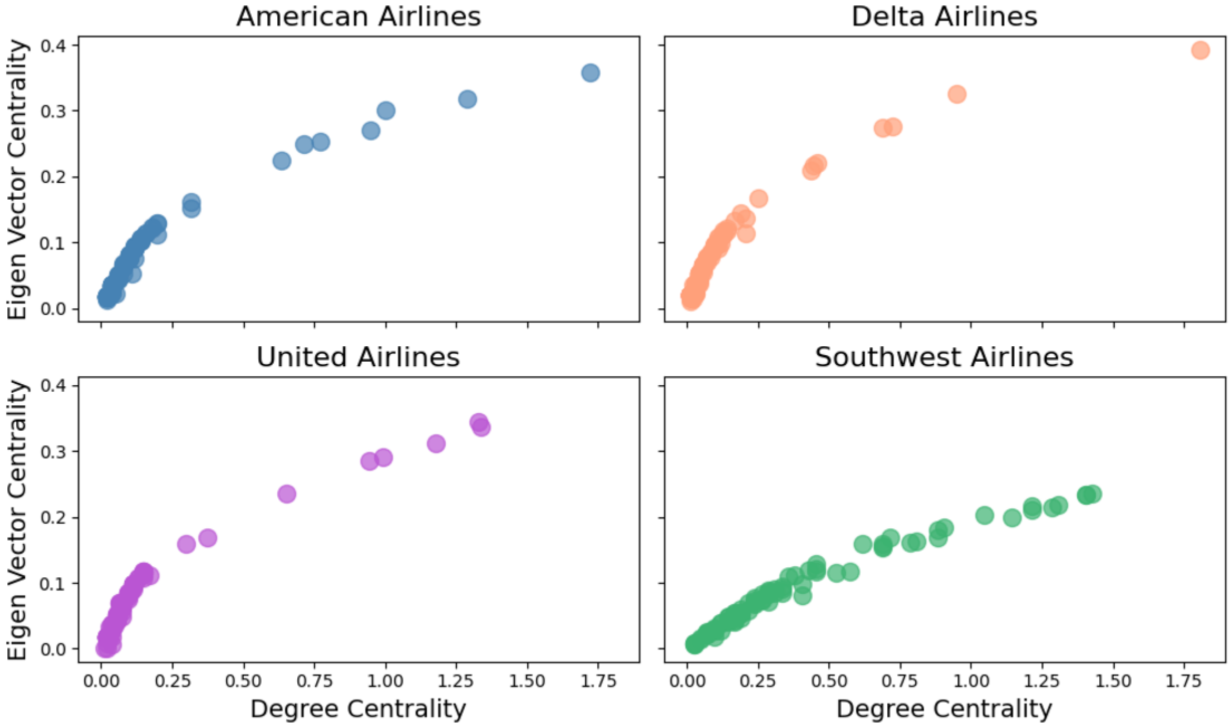
Much like betweenness centrality, the hubs of legacy carriers exhibit high values for both degree centrality and closeness centrality. Additionally, there are various sub or regional hubs associated with United and American Airlines that fall within the mid-range for degree and closeness values. Notably, it is interesting to observe that nearly all nodes across these airlines display at least mid-range closeness values, except for a specific node within United Airlines that records both degree and closeness values of zero.

This particular node is undoubtedly a spoke airport with minimal connections. Closeness provides insight into whether a node is part of an efficient pathway within the network. Generally, spokes in legacy carriers also demonstrate mid to high closeness values, as they are typically directly linked to major hubs, facilitating efficient routes. The anomaly observed in United Airlines may indicate an airport connected to nodes that are not hubs.

Furthermore, it's worth noting that no nodes in Southwest Airlines exhibit low closeness values, suggesting that every airport is connected to others via efficient pathways, even if they are linked to fewer nodes.

The following presents a series of four scatter plots organized by airline, illustrating the relationship between degree centrality (on the x-axis) and eigenvector centrality (on the y-axis).

Figure 12: Comparison plot of Airlines on Eigen Vector Centrality



The eigenvector centrality provides insight into the significance of a node's connections. When a node is linked to other important nodes, it tends to receive a high eigenvector value; conversely, if it connects to less significant nodes, its value decreases. For legacy carriers, both hub airports and regional hubs exhibit high eigenvector values. Notably, the highest eigenvector value for United Airlines is quite similar to those of other legacy carriers, which contrasts with the variations seen in other centrality measures. This is attributed to the metric's focus on the influence of nodes as well as the quality of their connections. In contrast, spoke airports within legacy carriers display lower eigenvector values due to their diminished influence.

For Southwest Airlines, it is important to note that eigenvector values are lower compared to those of other airlines, which can be attributed to the different network structure employed. However, it should also be highlighted that nodes with high degree centrality in Southwest Airlines correspondingly have high eigenvector values. This may be due to the interconnected nature of these nodes, enhancing their influence relative to airports that exhibit less interconnectivity.

VI. Conclusions

This study examines airline network structures, focusing on the trade-offs between the Hub-and-Spoke and Point-to-Point models. By analyzing key performance metrics such as delays, load factors, and stage lengths among legacy and low-cost carriers, the research reveals the operational and economic dynamics of these systems.

The findings show that while the Hub-and-Spoke model provides superior connectivity and economies of scale, it is vulnerable to disruptions due to its reliance on centralized hubs, making it suitable for legacy carriers like American Airlines, Delta Air Lines, and United Airlines. In contrast, the Point-to-Point model, exemplified by Southwest Airlines, offers shorter travel times and flexibility, appealing to passengers seeking direct routes and low fares.

The analysis highlights how airlines are evolving to stay competitive. Legacy carriers are adopting features from low-cost airlines to attract budget-conscious travelers, while low-cost carriers are expanding into underserved markets. Loyalty programs, technological advancements, and targeted service expansions demonstrate the adaptability of both models in a changing aviation landscape.

Going forward, airlines must balance operational efficiency with passenger satisfaction. Legacy carriers should focus on optimizing hub operations, while low-cost carriers need to maintain flexibility and cost efficiency to enhance their market share. This research offers insights for refining airline networks and improving overall industry resilience.

VII. References

Baladram, S. (2024, October 9). *Decision Tree Regressor Explained: A Visual Guide with Code Examples*. Retrieved from Towards Data Science: https://towardsdatascience.com/decision-tree-regressor-explained-a-visual-guide-with-code-examples-fbd2836c3bef

Borenstein, S. (1989). Hub and high fares: Dominance and market power in the U.S. airline industry. *RAND Journal of Economics, 20(3)*, 344-365.

Bureau of Transportation Statistics. (n.d.). *Air Carriers : T-100 Segment (US Carriers Only)*. Retrieved from Bureau of Transportation Statistics: https://www.transtats.bts.gov/DL_SelectFields.aspx?gnoyr_VQ=GDM&QO_fu146_anzr=Nv4%20Pn44vr45

Bureau of Transportation Statistics. (n.d.). *On-Time : Reporting Carrier On-Time Performance*. Retrieved from Bureau of Transportation Statistics: https://www.transtats.bts.gov/DL_SelectFields.aspx?gnoyr_VQ=FGJ&QO_fu146_anzr=b0-gvzr

Button, K. (2002). Airline network economics. *D. Jenkins (Ed.), Handbook of airline economics*, (2nd ed., pp. 27-33).

Cook, G. N., & Goodwin, J. (2008). Airline Networks: A Comparison of Hub-and-Spoke and Point-to-Point Systems. *Journal of Aviation/Aerospace Education & Research, 17(2)*, 51-60.

Donoghue, J. (2002, June). Hub machine. *Air Transport World, 39(6)*, 5.

Franke, M. (2004). Competition between network carriers and low-cost carriers -retreat battle or breakthrough to a new level of efficiency. *Journal of Air Transport Management, 10*, 15-21.

Gillen, D., & Morrison, G. (2005). Regulation, competition, and network evolution in aviation. *Journal of Air Transport Management, 11*, 161-174.

IBISWorld. (2024). *Domestic airlines in the US (Industry report 48111B).*

Li, L. (2019, May 15). *Classification and regression analysis with decision trees*. Retrieved from Towards Data Science: https://towardsdatascience.com/https-medium-com-lorrli-classification-and-regression-analysis-with-decision-trees-c43cdbc58054

Nasrollahi, M., & Kordani, A. A. (2023). Designing Airline Hub-and-Spoke Network and Fleet Size by a Bio Objective Model Based on Passenger Preferences and Value of time. *Journal of Advanced Transportation*, 21 pages.

NetworkX Developers. (n.d.). *Betweenness centrality*. Retrieved from NetworkX: https://networkx.org/documentation/stable/reference/algorithms/generated/networkx.algorithms.centrality.betweenness_centrality.html

NetworkX Developers. (n.d.). *Closeness centrality*. Retrieved from NetworkX: https://networkx.org/documentation/stable/reference/algorithms/generated/networkx.algorithms.centrality.closeness_centrality.html

NetworkX Developers. (n.d.). *Degree centrality*. Retrieved November 18, 2024, from NetworkX: https://networkx.org/documentation/stable/reference/algorithms/generated/networkx.algorithms.centrality.degree_centrality.html

NetworkX Developers. (n.d.). *Ego graph*. Retrieved from NetworkX: https://networkx.org/documentation/stable/reference/generated/networkx.generators.ego.ego_graph.html

NetworkX Developers. (n.d.). *Eigenvector centrality*. Retrieved from NetworkX: https://networkx.org/documentation/stable/reference/algorithms/generated/networkx.algorithms.centrality.eigenvector_centrality.html

Raschka, S. (n.d.). *Sequential feature selector*. Retrieved from MLxtend: https://rasbt.github.io/mlxtend/api_subpackages/mlxtend.feature_selection/#sequentialfeatureselector

Scikit-learn Developers. (n.d.). *Decision tree regressor*. Retrieved from Scikit-learn: https://scikit-learn.org/dev/modules/generated/sklearn.tree.DecisionTreeRegressor.html

Scikit-learn Developers. (n.d.). *Linear regression*. Retrieved from Scikit-learn: https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html

Statsmodels Developers. (n.d.). *Ordinary least squares*. Retrieved from Statsmodels: https://www.statsmodels.org/dev/generated/statsmodels.regression.linear_model.OLS.html

Appendices

Data Dictionary of Reporting Carrier On-Time Performance

YEAR	year
MONTH	month
OP_UNIQUE_CARRIER	Unique Carrier Code
OP_CARRIER_FL_NUM	Flight Number
ORIGIN_AIRPORT_ID	Origin Airport, Airport ID. An identification number assigned by US DOT to identify a unique airport.
ORIGIN	Origin Airport Code
ORIGIN_CITY_NAME	Origin Airport City Name
DEST_AIRPORT_ID	Destination Airport, Airport ID. An identification number assigned by US DOT to identify a unique airport
DEST	Destination Airport Code
DEST_CITY_NAME	Destination Airport City Name
CRS_DEP_TIME	Scheduled departure time (local time: hhmm)
DEP_TIME	Actual Departure Time (local time: hhmm)
DEP_DELAY	Delay in minutes for departure. Early departures show negative numbers.
DEP_DELAY_NEW	Difference in minutes between scheduled and actual departure time. Early departures set to 0.

DEP_DEL15	Departure Delay Indicator, 15 Minutes or More (1=Yes)
DEP_DELAY_GROUP	Departure Delay intervals, every (15 minutes from <-15 to >180)
DEP_TIME_BLK	CRS Departure Time Block, Hourly Intervals (Scheduled)
TAXI_OUT	The period of time that a flight spends moving on the ground from the departure gate to the point of takeoff on the runway.
WHEELS_OFF	Wheels Off Time (local time: hhmm)
WHEELS_ON	Wheels On Time (local time: hhmm)
TAXI_IN	The period of time that a flight spends moving on the ground from the landing at the runway to the arrival gate.
CRS_ARR_TIME	Scheduled Arrival Time (local time: hhmm)
ARR_TIME	Actual Arrival Time (local time: hhmm)
ARR_DELAY	Difference in minutes between scheduled and actual arrival time. Early arrivals show negative numbers.
ARR_DELAY_NEW	Difference in minutes between scheduled and actual arrival time. Early arrivals set to 0.
ARR_DEL15	Arrival Delay Indicator, 15 Minutes or More (1=Yes)

ARR_DELAY_GROUP	Arrival Delay intervals, every (15-minutes from <-15 to >180)
ARR_TIME_BLK	CRS Arrival Time Block, Hourly Intervals (Scheduled)
CANCELLED	Cancelled Flight Indicator (1=Yes)
CANCELLATION_CODE	Specifies The Reason For Cancellation
DIVERTED	Diverted Flight Indicator (1=Yes)
CRS_ELAPSED_TIME	CRS Elapsed Time of Flight, in Minutes
ACTUAL_ELAPSED_TIME	Elapsed Time of Flight, in Minute
AIR_TIME	Flight Time, in Minutes
FLIGHTS	Number of Flights (1)
DISTANCE	Distance between airports (miles)
DISTANCE_GROUP	Distance Intervals, every 250 Miles, for Flight Segment
CARRIER_DELAY	Carrier Delay, in Minutes
WEATHER_DELAY	Weather Delay, in Minutes
NAS_DELAY	National Air System Delay, in Minutes
SECURITY_DELAY	Security Delay, in Minutes
LATE_AIRCRAFT_DELAY	Late Aircraft Delay, in Minutes
FIRST_DEP_TIME	First Gate Departure Time at Origin Airport (local time: hhmm)

TOTAL_ADD_GTIME	(Total Additional Ground Time - the total amount of extra time a flight spends on the ground beyond what was initially planned) Total Ground Time Away from Gate for Gate Return or Cancelled Flight
LONGEST_ADD_GTIME	(Longest Additional Ground Time) Longest Time Away from Gate for Gate Return or Cancelled Flight
DIV_AIRPORT_LANDINGS	Number of Diverted Airport Landings
DIV_REACHED_DEST	Diverted Flight Reaching Scheduled Destination Indicator (1=Yes)
DIV_ACTUAL_ELAPSED_TIME	Elapsed Time of Diverted Flight Reaching Scheduled Destination, in Minutes. The ActualElapsedTime column remains NULL for all diverted flights.
DIV_ARR_DELAY	Difference in minutes between scheduled and actual arrival time for a diverted flight reaching scheduled destination. The ArrDelay column remains NULL for all diverted flights.
DIV_DISTANCE	Distance between scheduled destination and final diverted airport (miles). Value will be 0 for diverted flight reaching scheduled destination.

DIV1_AIRPORT	Diverted Airport Code1
DIV1_AIRPORT_ID	Airport ID of Diverted Airport 1. Airport ID is a Unique Key for an Airport
DIV1_WHEELS_ON	Wheels On Time (local time: hhmm) at Diverted Airport Code1
DIV1_TOTAL_GTIME	Total Ground Time Away from Gate at Diverted Airport Code1
DIV1_LONGEST_GTIME	Longest Ground Time Away from Gate at Diverted Airport Code1
DIV1_WHEELS_OFF	Wheels Off Time (local time: hhmm) at Diverted Airport Code1
DIV1_TAIL_NUM	Aircraft Tail Number for Diverted Airport Code1
DIV2_AIRPORT	Diverted Airport Code2
DIV2_AIRPORT_ID	Airport ID of Diverted Airport 2. Airport ID is a Unique Key for an Airport
DIV2_WHEELS_ON	Wheels On Time (local time: hhmm) at Diverted Airport Code2
DIV2_TOTAL_GTIME	Total Ground Time Away from Gate at Diverted Airport Code2
DIV2_LONGEST_GTIME	Longest Ground Time Away from Gate at Diverted Airport Code2

DIV2_WHEELS_OFF	Wheels Off Time (local time: hhmm) at Diverted Airport Code2
DIV2_TAIL_NUM	Aircraft Tail Number for Diverted Airport Code2

Cancellation Code

Code	**Description**
A	Carrier
B	Weather
C	National Air System
D	Security

DIV_AIRPORT_LANDINGS

Code	**Description**
0	Flight is not Diverted
1	One Diverted Airport Landing
2	Two Diverted Airport Landings
3	Three Diverted Airport Landings
4	Four Diverted Airport Landings
5	Five Diverted Airport Landings
9	Air Return to Origin Airport where the Flight was Ultimately Cancelled

Data Dictionary of T – 100 Segment (Traffic Data)

Year	Year
Month	Month
Unique_carrier	Unique Carrier Code
Unique_carrier_name	Unique Carrier Name
Data_source	Source of Data (D=Domestic ,I=International)
Class	Service Class
Aircraft_type	Aircraft Type
Origin_Airport_id	Origin Airport, Airport ID. An identification number assigned by US DOT to identify a unique airport.
Origin	Origin Airport Code
Origin_city_name	Origin Airport City Name
Dest_Airport_Id	Destination Airport, Airport ID. An identification number assigned by US DOT to identify a unique airport
Dest	Destination Airport Code
Dest_city_name	Destination Airport City Name
Distance_grouped	Ranges of miles into different groups
Departures_scheduled	Departures Scheduled
Departures_performed	Departures Performed
Payload	Total weight carried by aircraft (pounds)
Seats	Available Seats
Passengers	Non-Stop Segment Passengers Transported
Freight	Non-Stop Segment Freight Transported (pounds)

Mail	Non-Stop Segment Mail Transported (pounds)
Distance	Distance between airports (miles)
Ramp_to_Ramp	Ramp-to-Ramp Time, in Minutes
Air_Time	Air Time, in Minutes
ASM	Available Seats per Miles
RPM	Revenue Passenger Miles
Load_Factor	the proportion of airline output that is actually consumed
Stage_Length	The average distance flown, measure in statute miles, per aircraft departure
Block_Time Efficiency	time being spent in the air versus on the ground (taxiing in/out)