Forecasting Airfare Prices

By: Tony Lucci

AGENDA

Problem Statement

Airfare pricing Is a very interesting topic. Predicting the cost of airfare is a significant undertaking that a lot of individuals interested in travelling would benefit from. You know the price of a flight when you book of course. But wouldn't it be advantageous to know ahead of time what you will pay so when you are searching for your flight you have a grasp on if the price you are being quoted is a bargain?

There are many moving parts that go into the pricing of airfare:

- Price of Jet Fuel (Kerosine)
- Flight Distance
- Competition (imp comp: oligopoly, etc.)
- Timing of Purchase
- Timing of Flight
- Passenger Appetite (demand)
- Big Brother (security fees, taxes, etc.)
- Air Carrier Specific Fees
- Empty Middle Seats
- More!

Project objective

Create an ensemble of statistical **regression** models based on historical time series data (route specific) in order to **accurately predict airfare pricing by route**.

Time Series Models:

- *Univariate*
  - Ordinary Least Squares (OLS)
  - Auto Regressive Integrated Moving Average (ARIMA)

Measuring Success:

- R2
- RMSE

Data Collection

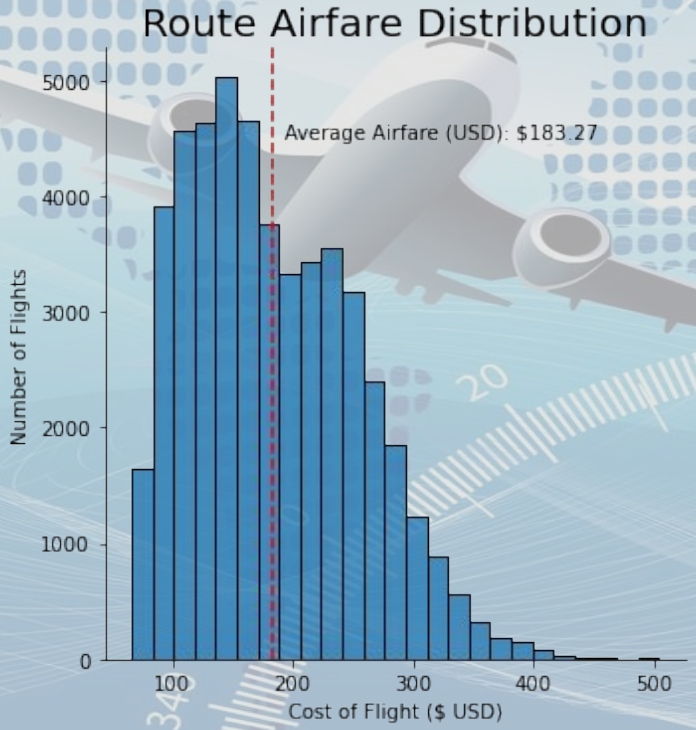
- **Historical Jet Fuel Prices**
  - Price of Jet Fuel in US Dollars.
  - Separated by month.
  - Ranges from April 1990 to August 2020.
- **Top 1,000 Contiguous State City-Pair Markets**
  - Average airfare prices per route by city for 48 landlocked US states
  - Separated by quarter.
  - Ranges from Q1 1996 to Q3 2019.
- US Domestic Flights
  - Flight data including route by city, route by airport, passengers, number of flights, total seats available, distance, origin population, and destination population
  - Separated by month.
  - Ranges from January 1990 to December 2009.
- Additional Data
  - City Latitude / Longitude

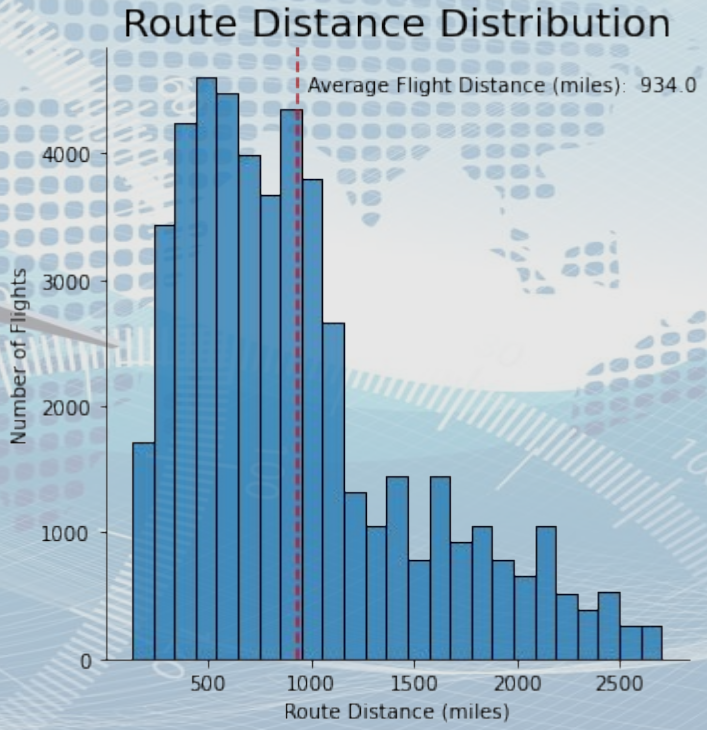
Post Data Merging and cleaning:

Combining the datasets together our final dataset contains...

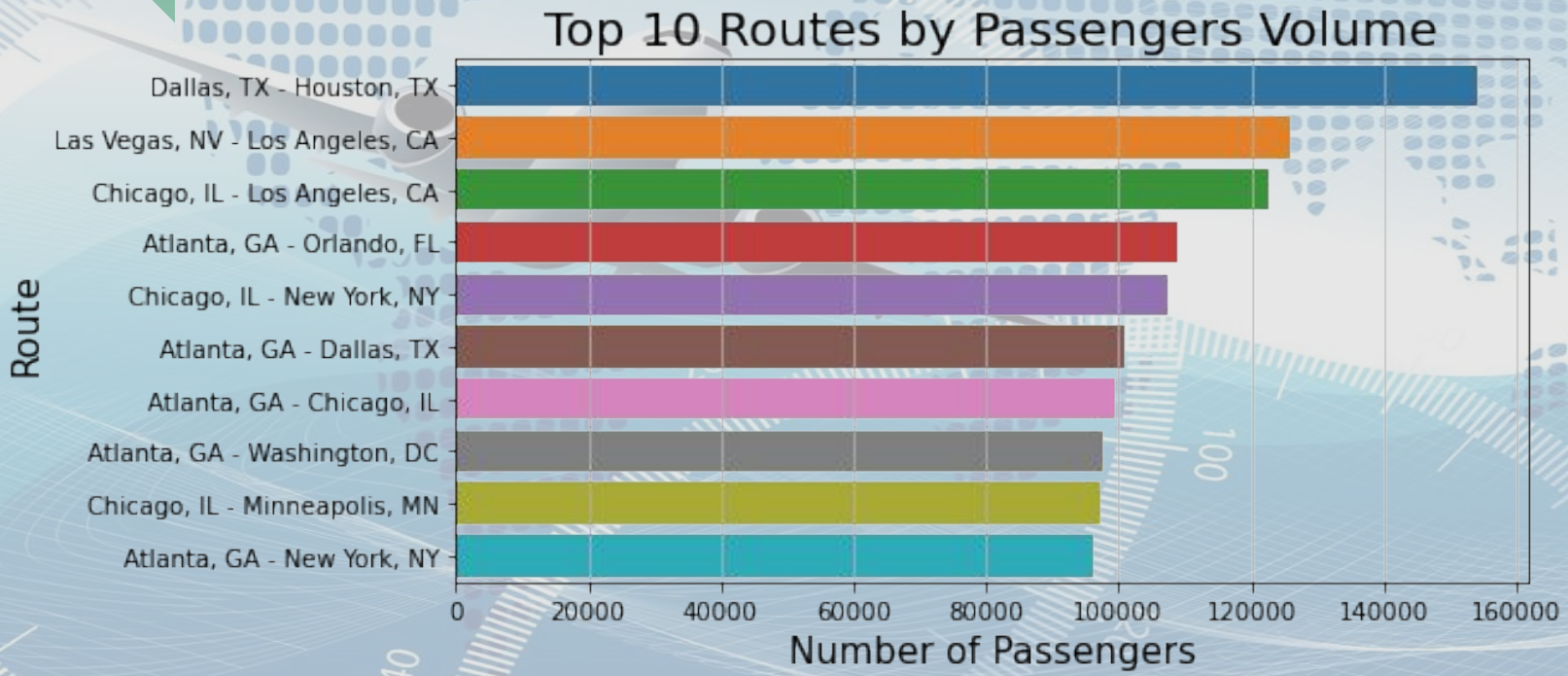
- 375 Routes
- 168 Months of Data
- 01/01/1996 - 12/31/2009
- Train/Test - 01/01/96 - 12/31/06
- Unseen - 01/01/07 - 12/31/09

Exploratory Data Analysis

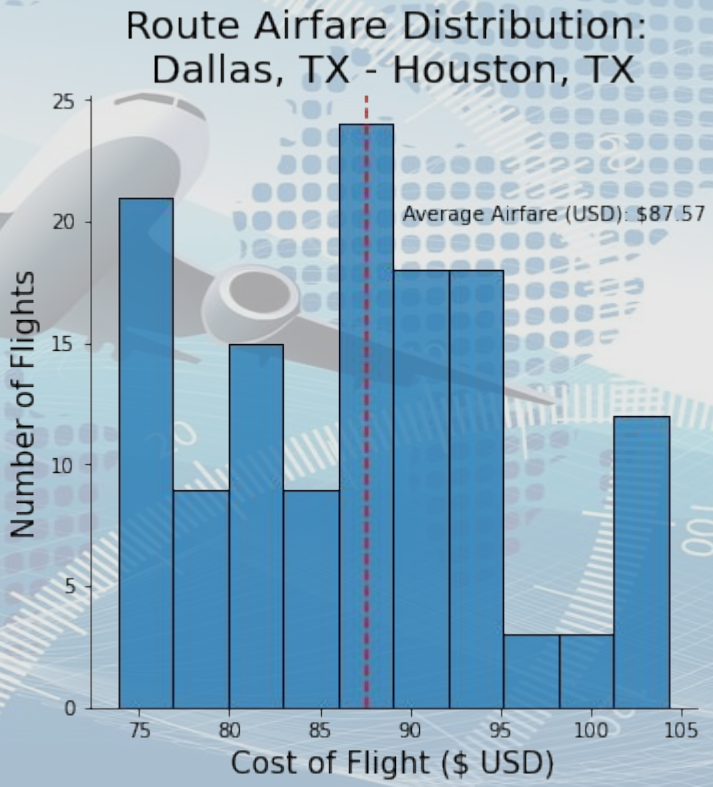


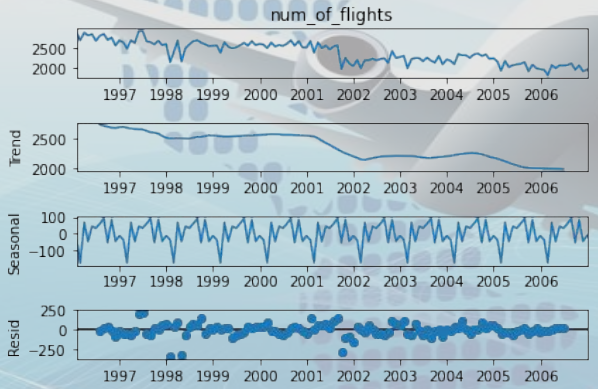


Exploratory Data Analysis



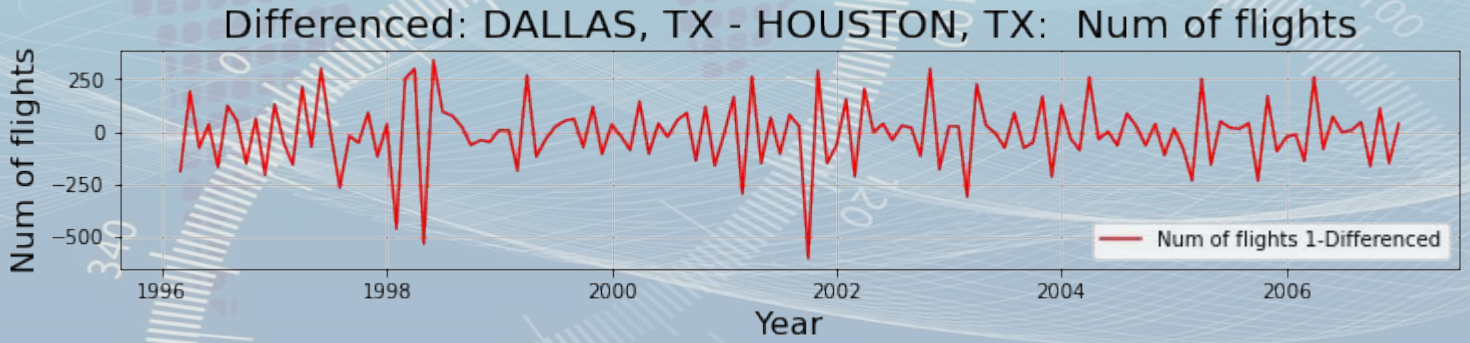
Exploratory Data Analysis

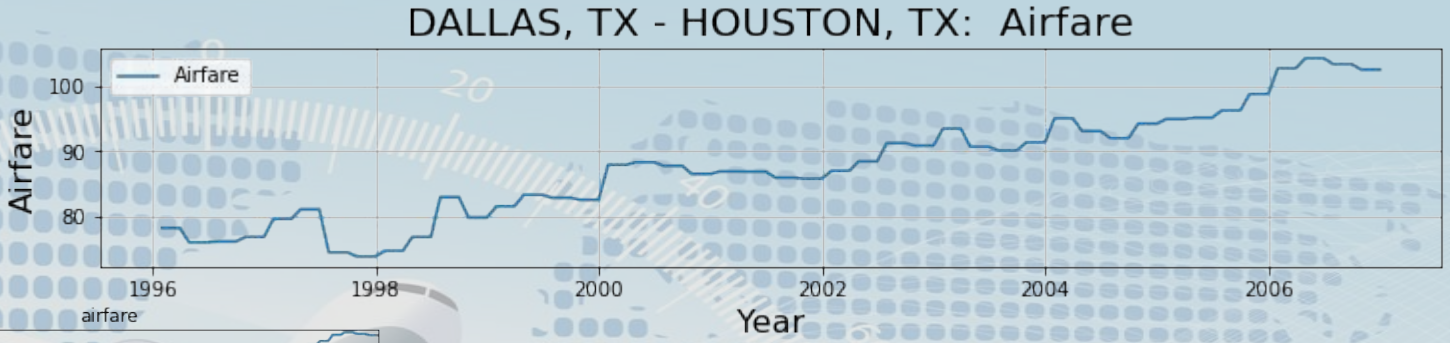


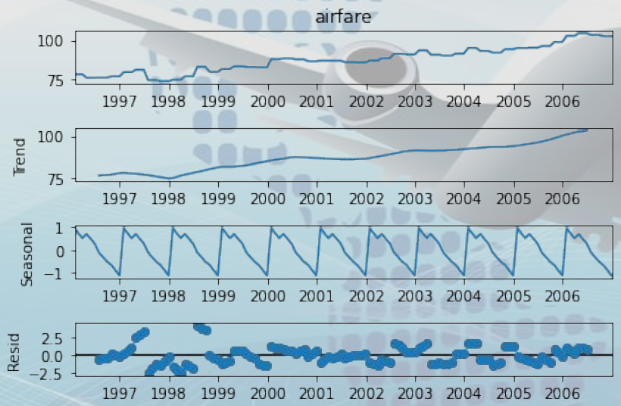


DATA IS NOT STATIONARY

CONDUCT ADFULLER TEST

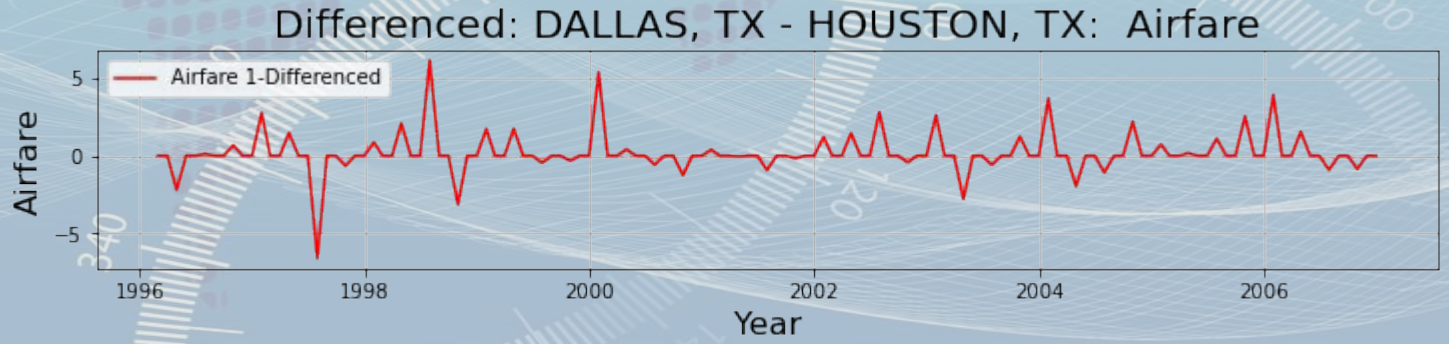


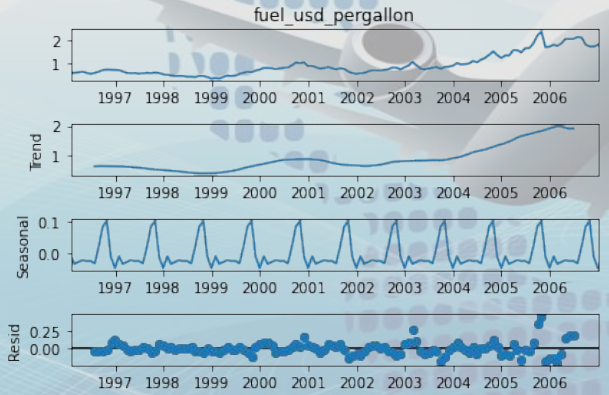




DATA IS NOT STATIONARY

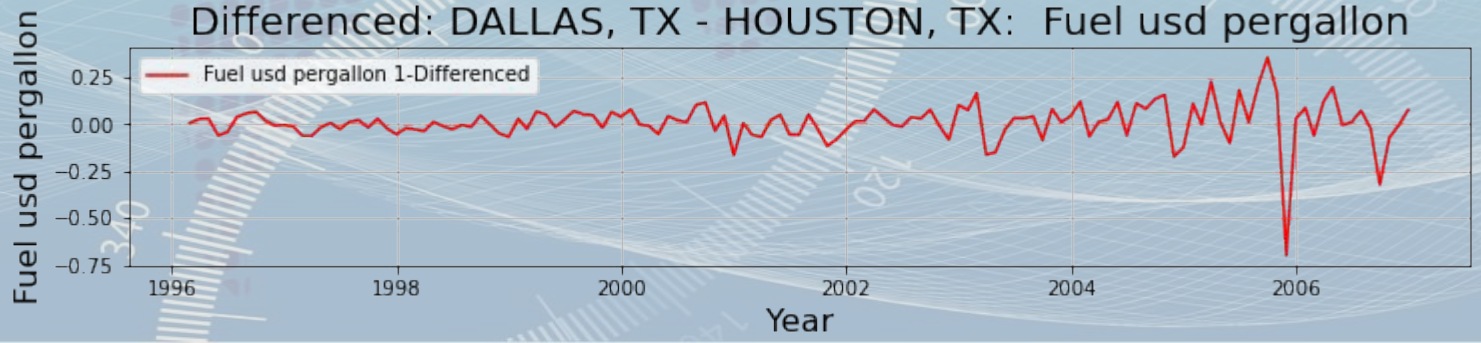
CONDUCT ADFULLER TEST

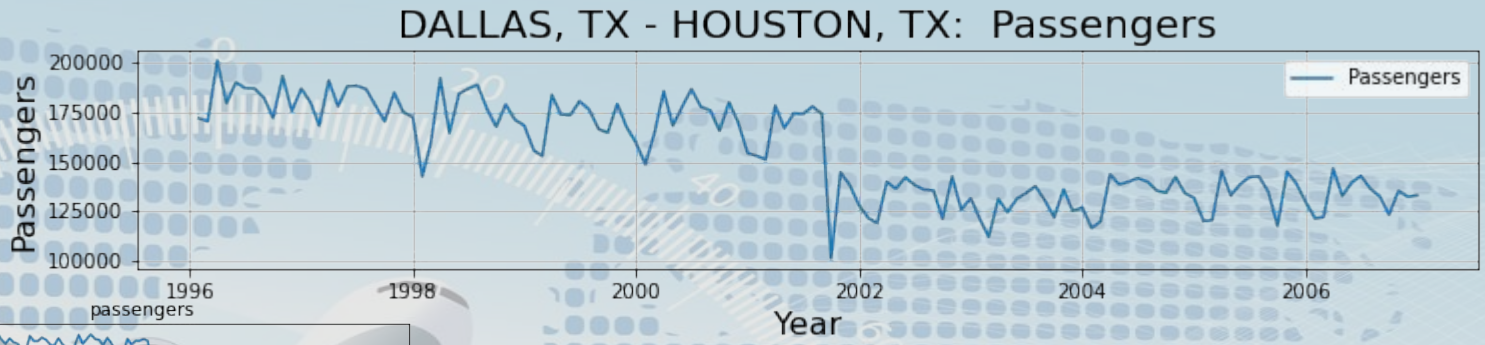


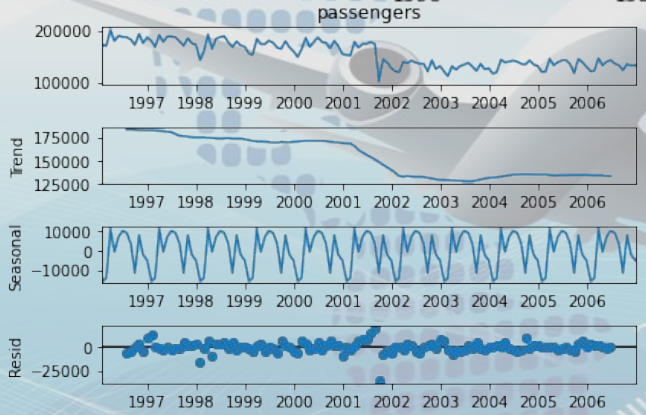


DATA IS NOT STATIONARY

CONDUCT ADFULLER TEST

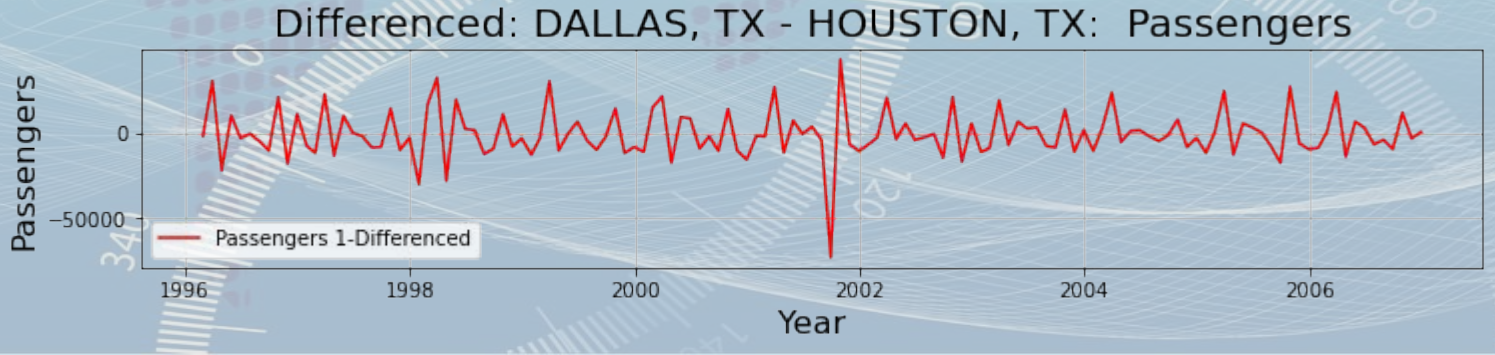


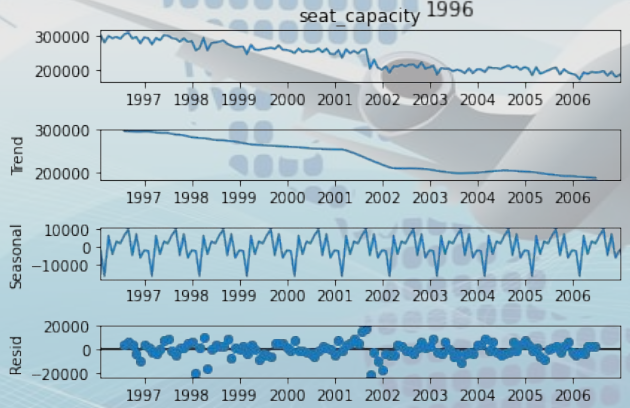




DATA IS NOT STATIONARY

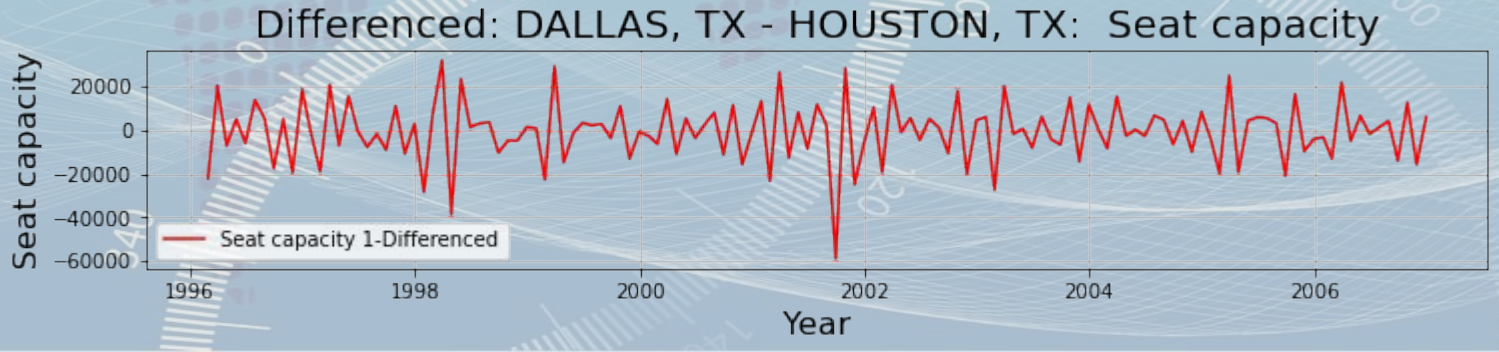
CONDUCT ADFULLER TEST





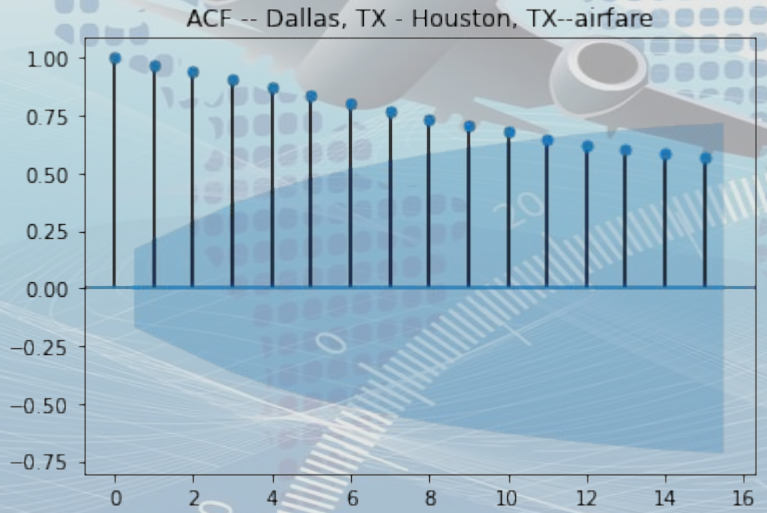
DATA IS NOT STATIONARY

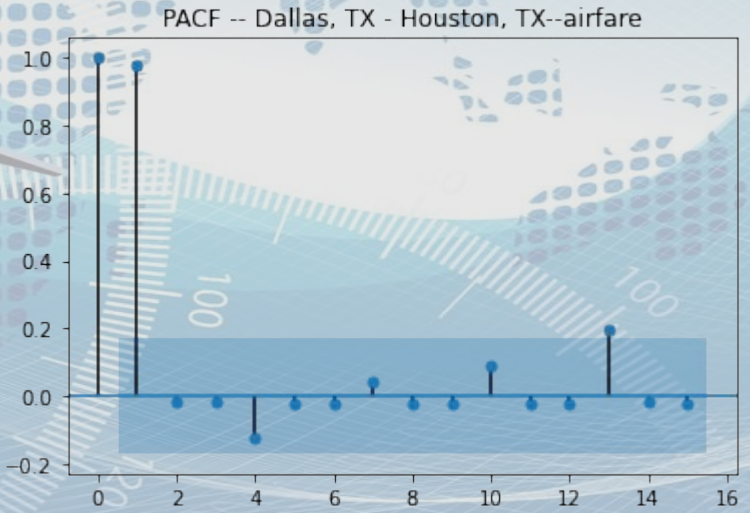
CONDUCT ADFULLER TEST



Exploratory Data Analysis

- **Auto-Correlation Plot & Partial Auto-Correlation Plot**
  - 95% confidence no trend (shaded blue)
  - Identify trends (ACF) and seasonality (ACF & PACF)
  - Engineer Lagged Features where correlations are identified
  - Best to analyze every feature!





Pre-Processing: Feature Engineering

Manually Engineered**

- Total Flight Miles
- Total Flight Cost
- Flight Demand
- Cost Per Mile
- Flight Revenue
- Passengers Per Flight
- Time (passed)

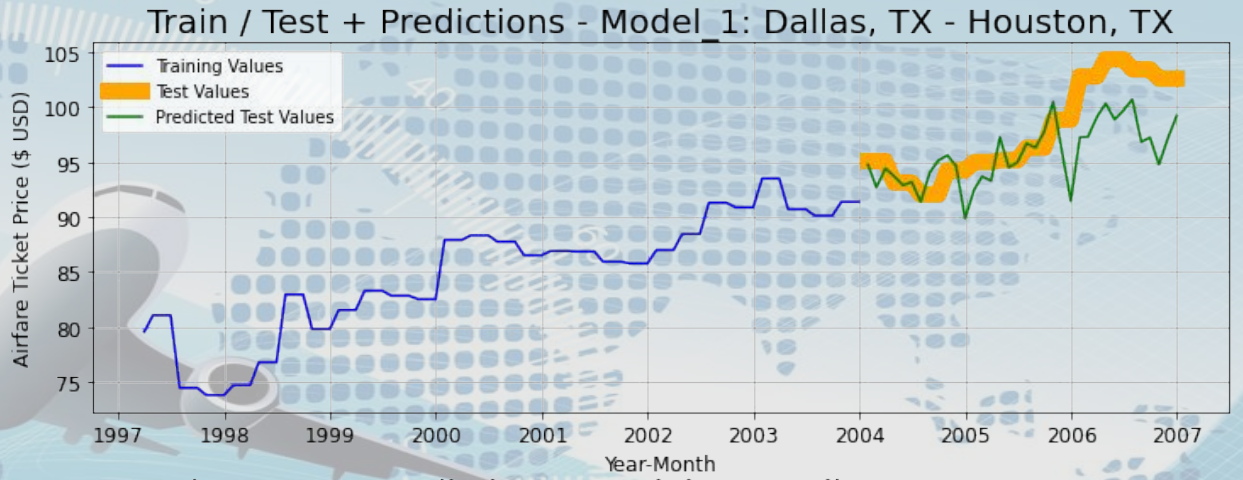
ACF / PACF Analysis Engineered**

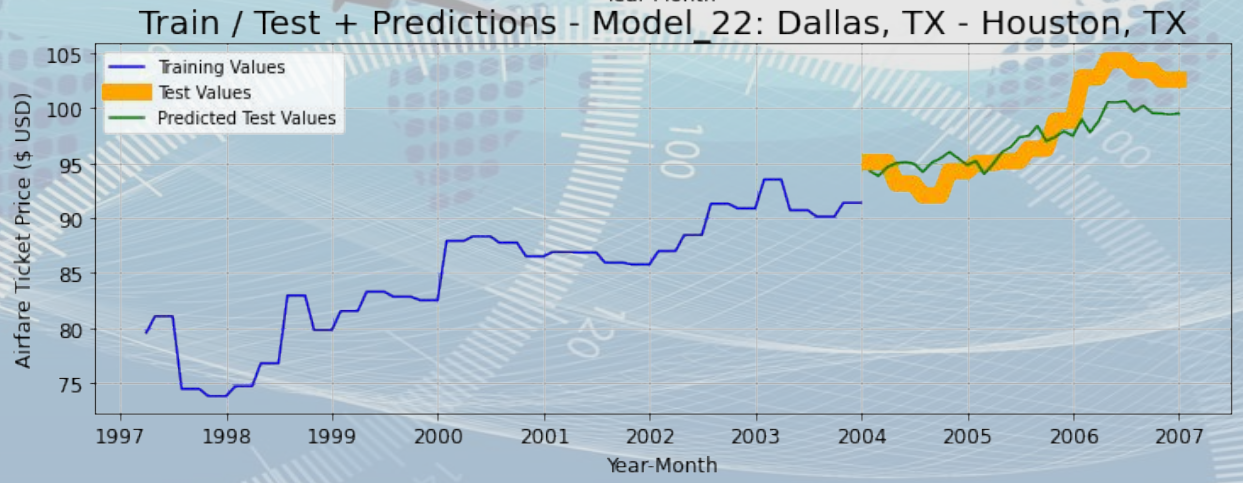
- Passengers 12
- Flight Demand 12
- Passengers Per Flight 12
- Flight Revenue 12
- Passengers 1
- Seat Capacity 1
- Airfare 1
- Fuel 1
- Flight Demand 1
- Passengers Per Flight 1
- Seat Capacity 2
- Number of Flights 2
- Number of Flights 3

***Different features engineered for each route*

Modeling: Ordinary Least Squares (OLS)

- Models: 22
- Initial Features - 27
- Ending Features - 6
- Benchmarks
  - Base R2 - 2.22^-16
  - *Base RMSE - $4.23*
- Initial Metrics
  - Train R2 - 87.85%
  - Test R2 - 28.77%
  - Train RMSE - $1.87
  - Test RMSE - $3.57
- Final Metrics
  - Train R2 - 85.62%
  - Test R2 - 64.59%
  - Train RMSE - $2.04
  - Test RMSE - $2.52
- Notes
  - AVG 2331 flights/month

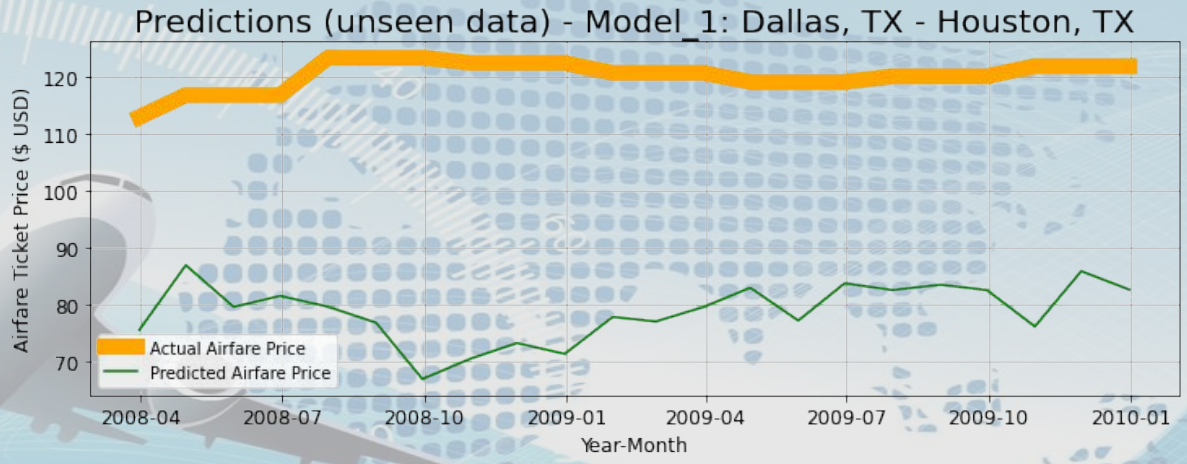


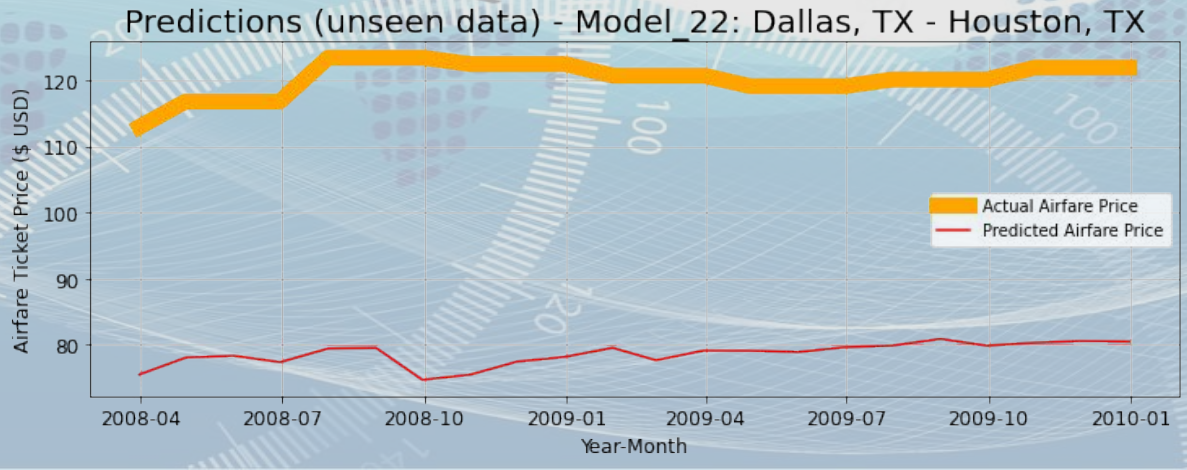


Modeling: Ordinary Least Squares (OLS)

- Final Features - 6
  - Number of Flights L3 D1
  - Flight Revenue D2
  - Number of Flights D1
  - Number of Flights L2 D1
  - Time
  - Passengers Per Flight L1 D2
- Scores
  - Unseen R2 - (260.82 * 100)
  - Unseen RMSE - $41.73

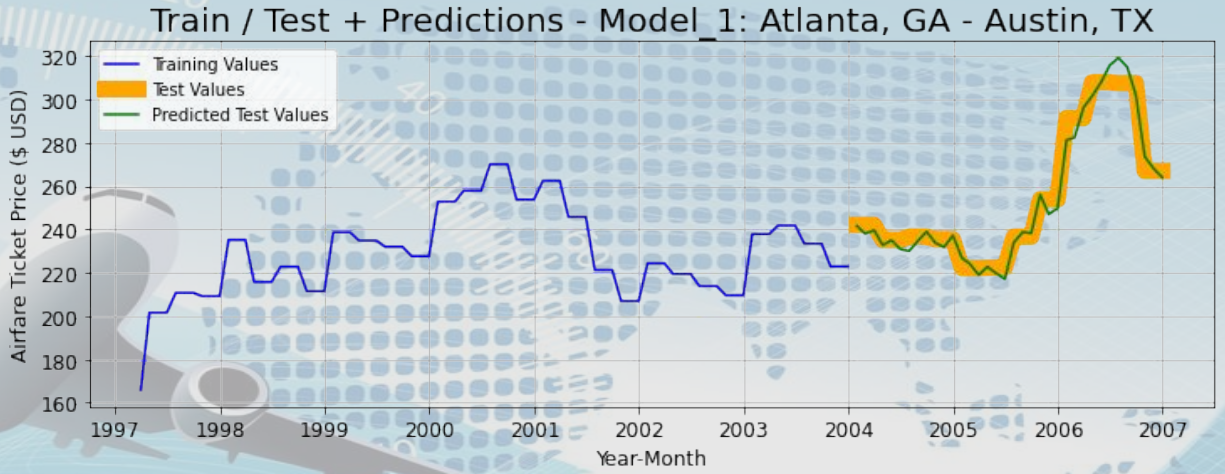
- **ARIMA**
  - AR (P) - 2 (Auto-Regressive)
  - I (D) - 1 (Difference/Stationary)
  - MA (Q) - 2 (Moving Average)
  - AIC (Akaike Information Criterion)
    - 363.74 (ARIMA)
    - 363.50 (OLS)

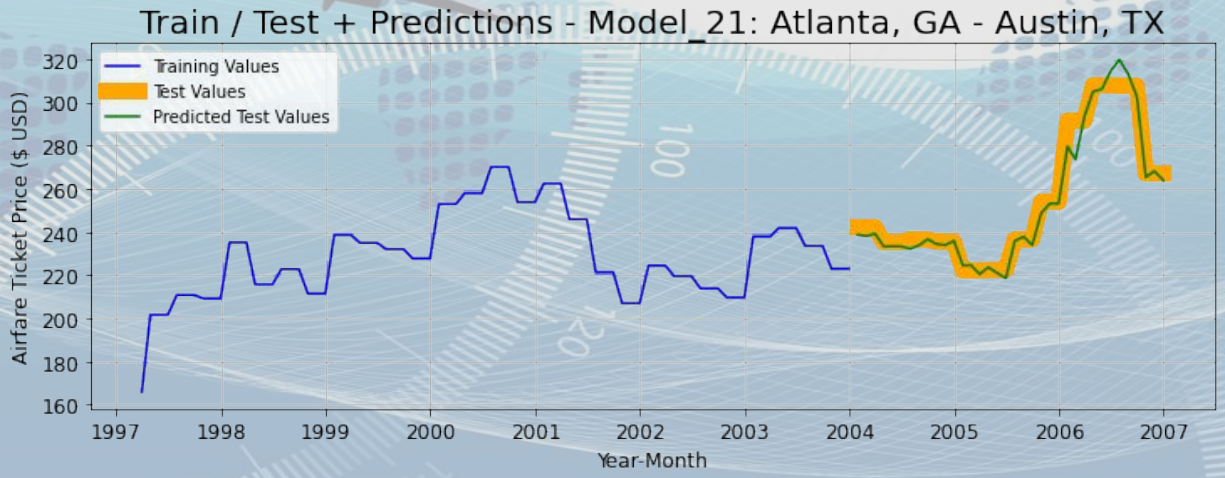




Modeling: Ordinary Least Squares (OLS)

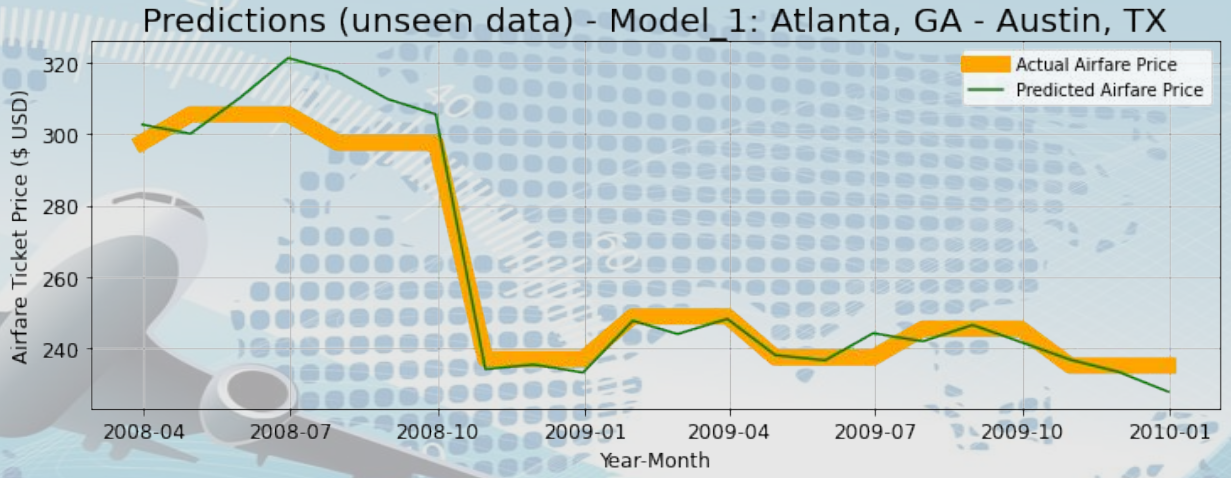
- Models: 21
- Initial Features - 27
- Ending Features - 7
- Benchmarks
  - *Base R2 - $0.00*
  - *Base RMSE - $29.99*
- Initial Metrics
  - Train R2 - 98.48%
  - Test R2 - 97.35%
  - Train RMSE - $2.38
  - Test RMSE - $4.89
- Final Metrics
  - Train R2 - 98.21%
  - Test R2 - 97.27%
  - Train RMSE - $2.58
  - Test RMSE - $4.95
- Notes
  - AVG 185 flights/month

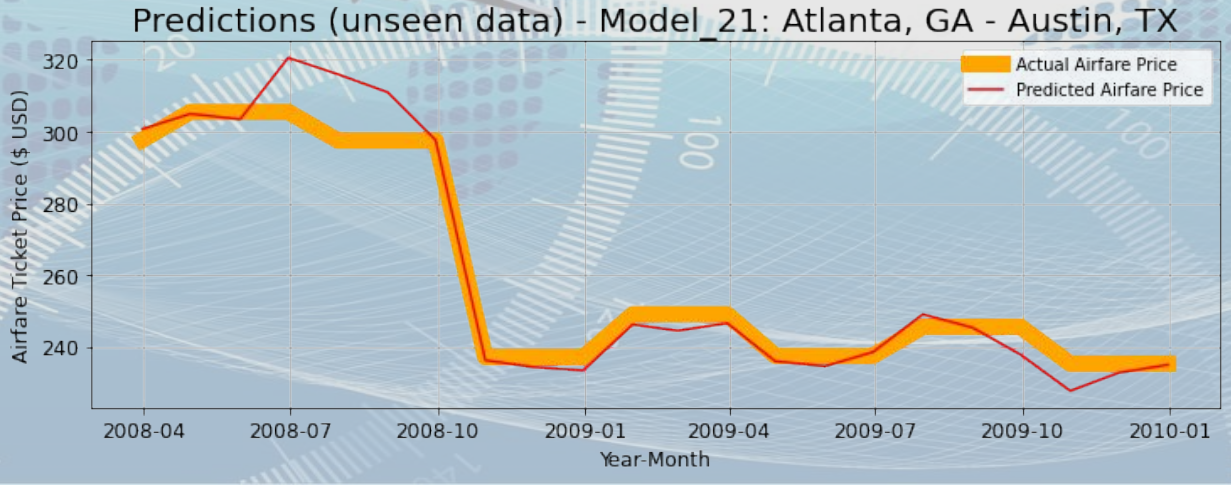




Modeling: Ordinary Least Squares (OLS)

- Final Features - 7
  - Total Flight Miles D1
  - Number of Flights L12 D2
  - Time D1
  - Seat Capacity L2
  - Passengers
  - Flight Distance
  - Flight Revenue
- Scores
  - Unseen R2 - 94.55%
  - Unseen RMSE - $6.64





Conclusions & Business Recommendations

- Able to reduce average RMSE by 21.85% -- Improved Score on 299 out of 375 routes
- Success with 63 out of 375 routes in predicting airfare prices near perfect r2
- Most Influential Features
  - Number of Passengers
  - Number of Flights
  - Flight Distance
- Predicting Airfare Pricing Is Very Tricky
  - Airlines are not quick to share their pricing strategies

FUTURE

- Obtain access to a top carriers daily pricing information + Scrape the Web for Hourly Information (peak/non-peak timing)
- Build Out Web App - Showcases Savings If you Buy Now VS. Future Pricing


QUESTIONS?