Rani Misra, Abel Garcia – Chavez, Alejandro Perez, Jade Reutzel

STA 4753: Project

Introduction

Airplanes are a key method of transportation across the vast regions of the United States. With 853 million passengers utilizing US airlines in 2022 alone, this form of transportation has been in the forefront of traveling within and outside the US ("Full Year 2022 U.S. Airline Traffic Data"). Traveling by air can be an enjoyable experience for some as it might be their first time leaving their state or the country, or they are just excited to visit a new destination. On the other hand, it can be extremely anxiety-inducing for others, especially when it comes to delaying or cancelling flights, overbooking seats, experiencing turbulence during the flight, or losing luggage during the journey.

Among multiple other concerns, lost luggage is a prevalent and deceitful issue as it does not present itself until after the passenger has completed a successful flight and is ready to head to their destination with their luggage, only for it to be missing. In 2022 alone, domestic flights supported by US airlines have reportedly lost over two and a half million pieces of luggage ("Baggage Mishandled by Marketing U.S. Air Carriers"). The typical solution to this concern has been reporting the lost luggage to the airline and hoping that they are able to find it. In most cases, passengers are entitled to monetary compensation for the contents of their luggage if they are left without their belongings for an unreasonable time. In recent times, passengers have chosen more technological methods to protect their luggage. While many airlines might not use these aids in their own pursuits, passengers have chosen to keep GPS trackers in their luggage like Tile or AirTags to keep an eye on their belongings and help airlines locate their baggage in case it is lost.

While consumers have found new tools to aid them in this predicament, passengers in the past were not privy to such technologies yet still faced this issue. This analysis focuses on identifying and forecasting the amount of reported lost luggage from two US airlines before the age of modern commercialized tracking tools to understand the trends present in the airline industry during this time. The two airlines to consider in this analysis are United and American Eagle.

United Airlines commenced operations in 1931 ("The Boeing Logbook: 1927–1932") and continues to operate an extensive domestic and international route network across the United States and all six inhabited continents primarily out of its eight hubs ("Star Alliance Facts and Figures"). They consistently rank as one of the world's largest airlines, ranking first by the number of destinations served and third in terms of revenue and fleet size. American Eagle Airlines was a subsidiary of the American Airlines Group that was operational from 1998 to 2014. During its operational years, the airline operated more than 1,000 daily flights to over 150 destinations in the United States, Canada, Mexico, Caribbean and South America. In early 2014, the airline rebranded to 'Envoy Air' to avoid confusion when American Airlines announced that other regional carriers would operate on behalf of American ("Envoy Air").

Discussion of the Data Set

The FAA requires air carriers to report flight delays, cancellations, overbookings, late arrivals, baggage complaints, and other operating statistics to the U.S. government, which compiles the data and reports it to the public. This data has been collected from government reports, compiled, and housed on Kaggle. Specifically, this data set contains monthly observations from 2004 to 2010 for United Airlines, American Eagle, and Hawaiian Airlines. The variables included are the name of the airline, the month and year of the report, baggage, scheduled, canceled, and enplaned. We provide a more elaborate explanation of the variables below:

- Baggage - The total number of passenger complaints for theft of baggage contents, or for lost, damaged, or misrouted luggage for the airline that month.
- Scheduled - The total number of flights scheduled by that airline that month.
- Canceled - The total number of flights canceled by that airline that month.
- Enplaned - The total number of passengers who boarded a plane with the airline that month.

We excluded Hawaiian Airlines from the analysis due to several considerations. As an airline devoted to routes from and to Hawaii to customers in Asia, American Samoa, Australia, French Polynesia, Hawaii, New Zealand, and the United States mainland, the overall operations is significantly smaller when compared to United Airlines and American Eagle Airlines ("Hawaiian Airlines."). In regard to fleet size during the period in which this data was collected, Hawaiian Airlines only had an average of around 30 operational aircrafts ("Hawaiian Airlines Fleet Details | Airfleets Aviation."), compared to United's 208 aircrafts ("United Airlines Fleet Details | Airfleets Aviation.") and American Eagle's 310 aircrafts ("American Eagle Airlines Fleet Details | Airfleets Aviation."). Thus, including them in the analysis would have skewed the data considering its peers' fleet sizes. Additionally, Hawaiian Airlines frequently tops the on-time carrier list in the United States, as well as the fewest cancellations, oversales, and baggage handling issues ("Hawaiian Airlines."), thereby further skewing the data which is focused on highlighting these key incidents. Focusing on national conglomerates allows us to examine trends that have a more substantial impact on the airline industry. By integrating these insights into the analysis, we gain a comprehensive understanding of trends, seasonality, and operational factors influencing lost baggage complaints within the US airline industry during the specified period.

Modeling

Analyzing the data from 2004 to 2010 reveals interesting trends in lost baggage complaints across United Airlines and American Eagle. A visual examination of the data shows a gradual decrease in reported lost baggage incidents over this period. This trend suggests improvements in baggage handling processes or passenger awareness regarding luggage protection. Seasonal patterns play a significant role in lost baggage complaints. The analysis indicates noticeable peaks in lost baggage incidents during the summer travel season and around winter holidays every year. These periods are associated with increased passenger traffic and potentially more complex logistics, contributing to higher rates of lost or mishandled luggage. The year 2007 stands out with a notable peak in lost baggage complaints. We may attribute this peak to several factors, including increased airline delays, changes in fleet operations, or other industry-specific challenges prevalent during that time. Simple research about the industry climate during this time corroborates this observation. More than 26 percent of commercial flights in the US arrived late or were canceled in 2007 as rising passenger demand and an industry preference for smaller planes intensified congestion in the skies and on runways (Caterinicchia). We could have attributed an increase in delays and cancellations to the peak travel seasons in damaged or lost baggage claims during this time, thus highlighting the impact of operational disruptions on baggage handling efficiency.

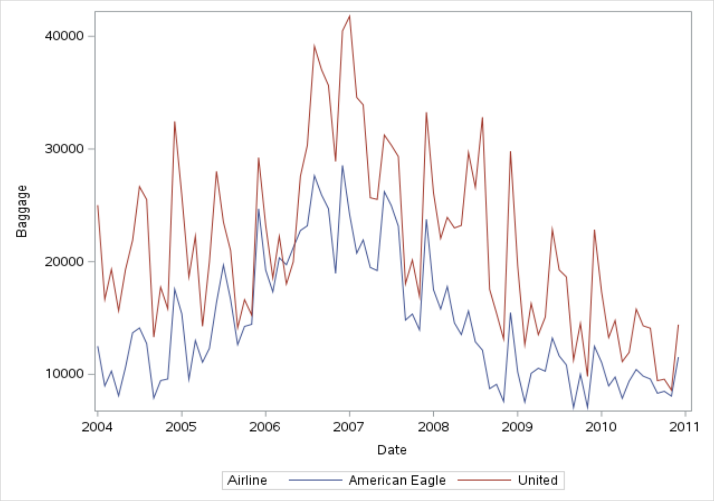


Figure 1: Number of baggage complaints from 2004-2010 for United and American Eagle Airlines

To further explore the data, differencing techniques were applied to address the downward trend observed. Differencing involves computing the difference between consecutive observations, which can help remove trends and make the data more stationary for modeling purposes. This step is crucial in ensuring that our analysis captures meaningful patterns without being skewed by long-term trends. We apply the first difference of the variable 'Baggage' to account for these factors. The result yields data that is much more stationary but still has a large variance.

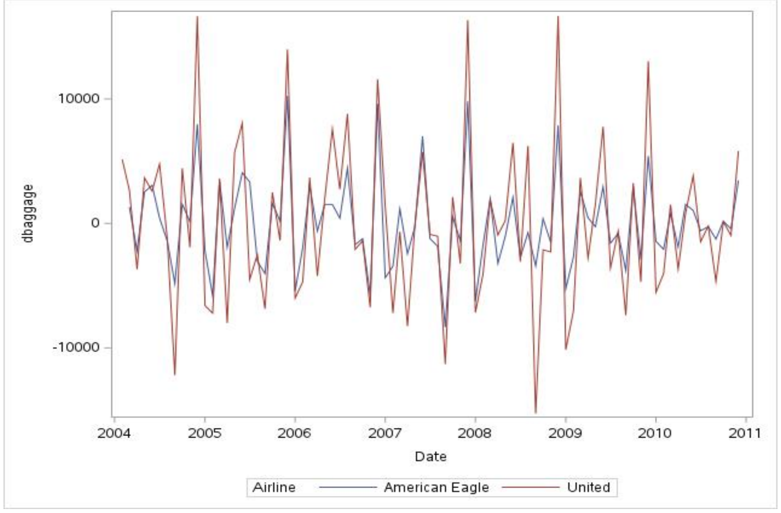


Figure 2: The first difference of Baggage

Thus, we take the cube root of the first difference of Baggage. This results in a range of variance that is much smaller and ideal for modeling.

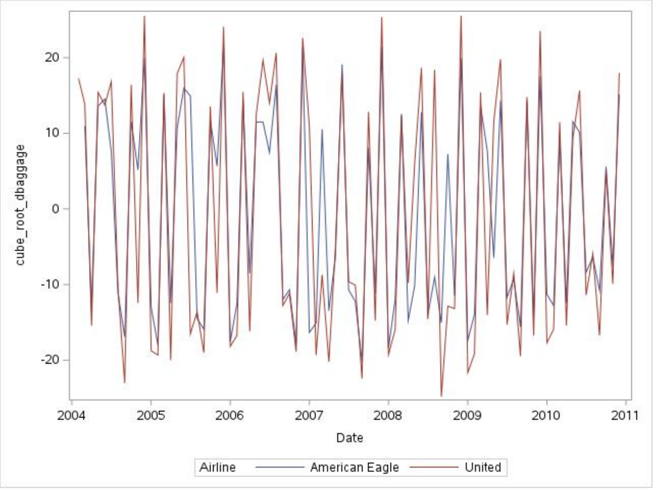


Figure 3: The cube root of the first difference of Baggage

We included a cross correlation function between the cube root of the first difference of Baggage (referred to as cube_root_baggage) and both the airlines to account for the possible industry trend or standard set by either airline regarding increased baggage complaints.

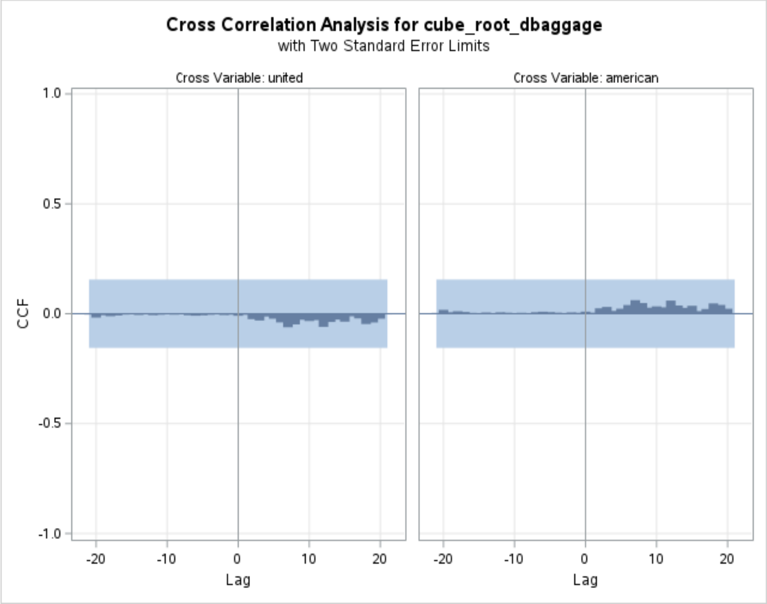
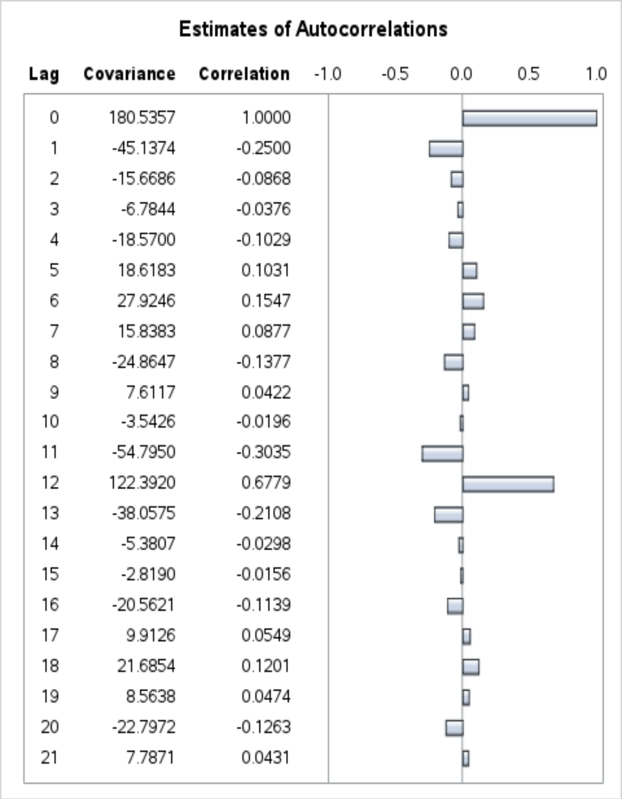


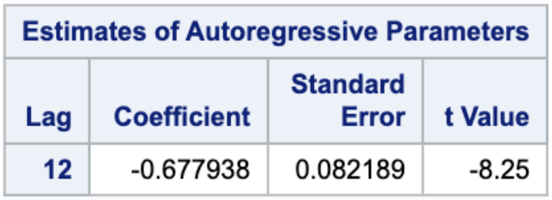
Figure 4: Cross Correlation Analysis for cube root of first difference with United and American Eagle Airlines

Using an AUTOREG model for American Eagle Airlines yielded a model with an AR(1,12) lag and an MA(12) lag obtained through backward elimination of autoregressive terms, as seen in Figures 5 and 6. Using an ARIMA model with the SCAN feature suggested an AR(1) or an MA(1) lag (p = 1 or q = 1) as seen in Figure 7.



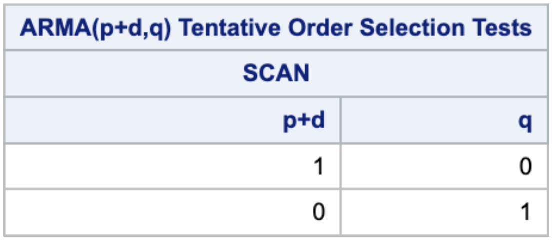
Estimates of Autocorrelations

Lag	Covariance	Correlation
0	180.5357	1.0000
1	-45.1374	-0.2500
2	-15.6686	-0.0868
3	-6.7844	-0.0376
4	-18.5700	-0.1029
5	18.6183	0.1031
6	27.9246	0.1547
7	15.8383	0.0877
8	-24.8647	-0.1377
9	7.6117	0.0422
10	-3.5426	-0.0196
11	-54.7950	-0.3035
12	122.3920	0.6779
13	-38.0575	-0.2108
14	-5.3807	-0.0298
15	-2.8190	-0.0156
16	-20.5621	-0.1139
17	9.9126	0.0549
18	21.6854	0.1201
19	8.5638	0.0474
20	-22.7972	-0.1263
21	7.7871	0.0431



Estimates of Autoregressive Parameters

Lag	Coefficient	Standard Error	t Value
12	-0.677938	0.082189	-8.25



ARMA(p+d,q) Tentative Order Selection Tests

SCAN

p+d	q
1	0
0	1

Figures 5 and 6: SAS Outputs showing AR lags (left) and MA lags (top right)

Figure 7: Suggested AR and MA lags using ARIMA SCAN (bottom right)

To choose the best model, we chose optimal values to maximize the AIC. The AUTOREG model has an AIC of 606.7476 whereas the ARIMA SCAN model has an AIC of 657.38. Thus, we analyze the AUTOREG model with an AR(1,12) lag and an MA(12) lag further. Figure 8 illustrates the various analytical thresholds the model must meet to be a good fit to the data. The ACF plot reinforces that we should include an MA(12) term in the model, and the PACF plot indicates that we should include an AR(11,12) lag. Analyzing a model that includes all these lags indicates that MA(12) and AR(11,12) are the only significant lags as they have an absolute value of the t value that is greater than one (Figure 9). This model results in a marginally improved AIC of 604.2482. In Figure 10, the residual diagnostics for this model meets all the required thresholds. However, the White Noise Probabilities could be concerning. It can be assumed that these probabilities will fall under 0.05 if additional differencing were to be applied.

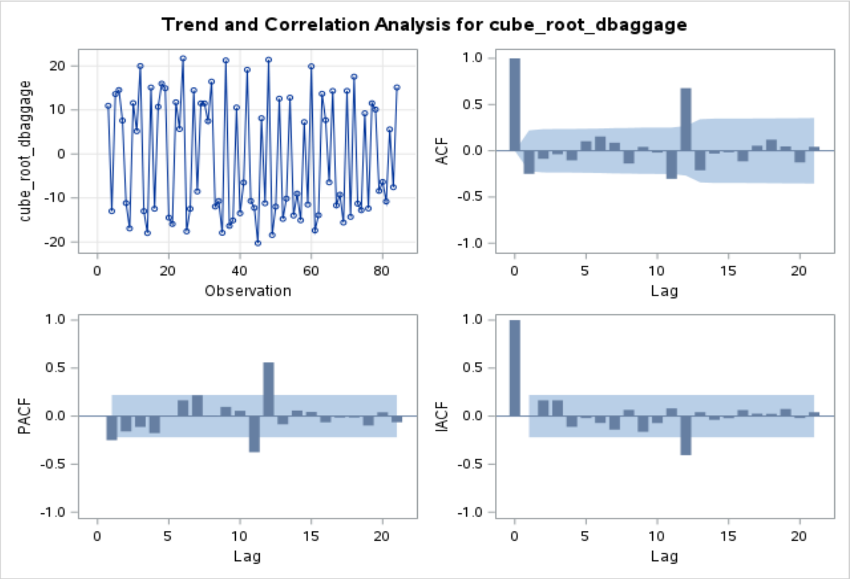
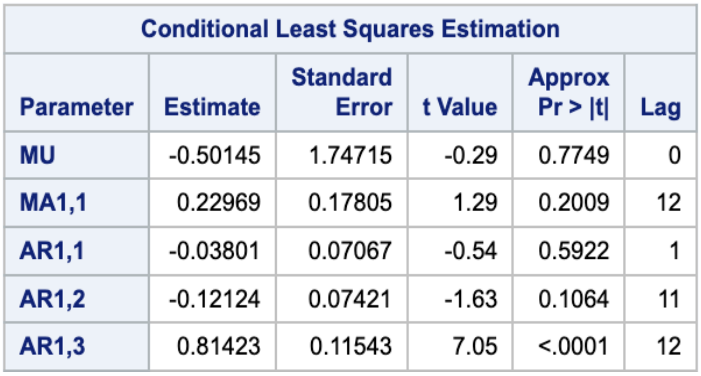


Figure 8: Trend and Correlation Analysis for P = (1 12) and Q = (12)

Figure 9: Significance of lags using P = (1 11 12) and Q = (12)



Conditional Least Squares Estimation					
Parameter	**Estimate**	**Standard Error**	**t Value**	**Approx Pr > |t|**	**Lag**
MU	-0.50145	1.74715	-0.29	0.7749	0
MA1,1	0.22969	0.17805	1.29	0.2009	12
AR1,1	-0.03801	0.07067	-0.54	0.5922	1
AR1,2	-0.12124	0.07421	-1.63	0.1064	11
AR1,3	0.81423	0.11543	7.05	<.0001	12

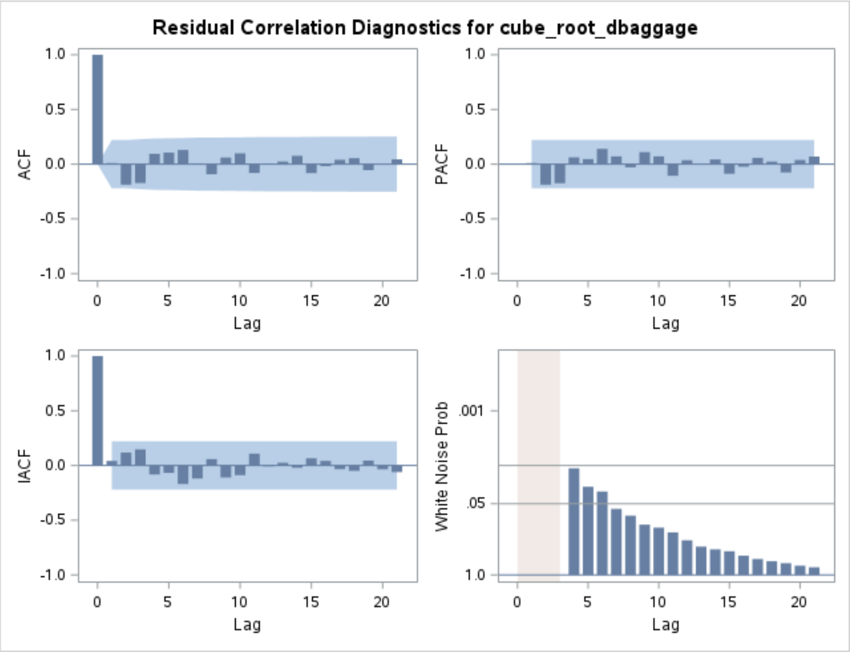
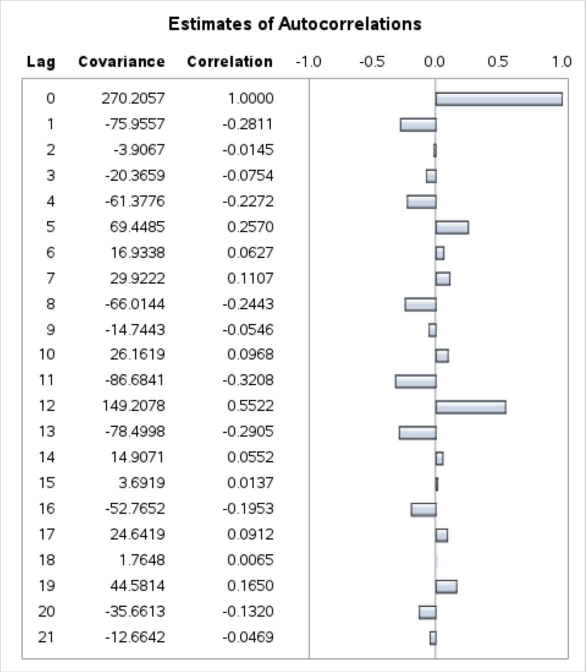


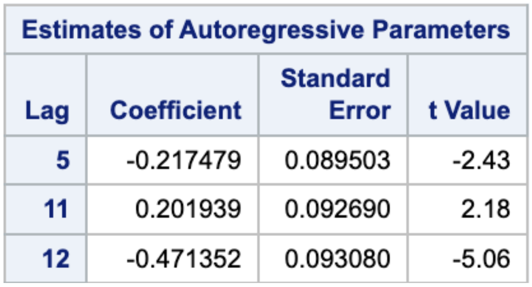
Figure 10: Residual Diagnostics for P = (11 12) and Q = (12)

Similarly, the AUTOREG model for United Airlines yielded a model with an AR(1,12) lag and an MA(5,11,12) lag obtained through backward elimination of autoregressive terms, as seen in Figures 11 and 12. Using an ARIMA model with the SCAN feature suggested an AR(2) lag or an MA(2) lag (p = 2 and q = 2) as seen in Figure 14.



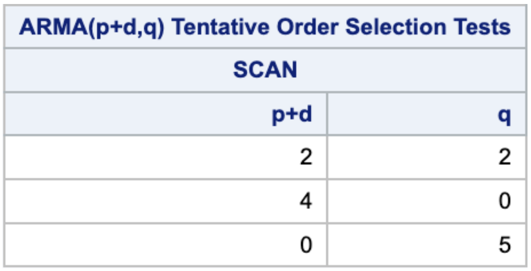
Estimates of Autocorrelations

Lag	Covariance	Correlation
0	270.2057	1.0000
1	-75.9557	-0.2811
2	-3.9067	-0.0145
3	-20.3659	-0.0754
4	-61.3776	-0.2272
5	69.4485	0.2570
6	16.9338	0.0627
7	29.9222	0.1107
8	-66.0144	-0.2443
9	-14.7443	-0.0546
10	26.1619	0.0968
11	-86.6841	-0.3208
12	149.2078	0.5522
13	-78.4998	-0.2905
14	14.9071	0.0552
15	3.6919	0.0137
16	-52.7652	-0.1953
17	24.6419	0.0912
18	1.7648	0.0065
19	44.5814	0.1650
20	-35.6613	-0.1320
21	-12.6642	-0.0469



Estimates of Autoregressive Parameters

Lag	Coefficient	Standard Error	t Value
5	-0.217479	0.089503	-2.43
11	0.201939	0.092690	2.18
12	-0.471352	0.093080	-5.06



ARMA(p+d,q) Tentative Order Selection Tests

SCAN

p+d	q
2	2
4	0
0	5

Figures 11 and 12: SAS Outputs showing AR lags (left) and MA lags (top right)

Figure 13: Suggested AR and MA lags using ARIMA SCAN (bottom right)

The AUTOREG model has an AIC of 655.7156 whereas the ARIMA SCAN model has an AIC of 699.7862. Thus, we further analyze the AUTOREG model with an AR(1,12) lag and an MA(5,11,12) lag. Figure 14 illustrates the various analytical thresholds the model must meet to be a good fit to the data. The ACF plot reinforces that we should include an MA(12) lag in the model but not MA(5) or MA(11), and the PACF plot indicates that we should also include an AR(11,12) lag. Analyzing a model that includes all these lags indicates that MA(12), AR(11), and AR(12) are the only significant lags as they have an absolute value of the t value that is greater than one (Figure 15). This model results in a marginally worse AIC of 658.6918. Reintroducing an MA(5) in the model is significant and improves the AIC to 656.5241. On the other hand, reintroducing an MA(11) in the model is insignificant and worsens the AIC to 672.176. Including both MA(5,11) and MA(12) is even worse off with an AIC of 686.2179. Thus, we will analyze the model with MA(5,12) and AR(11,12) lags further. In Figure 16, the residual diagnostics for this model meets all the required thresholds. However, the White Noise Probabilities could be concerning. We assume that these probabilities will fall under 0.05 if we apply additional differencing.

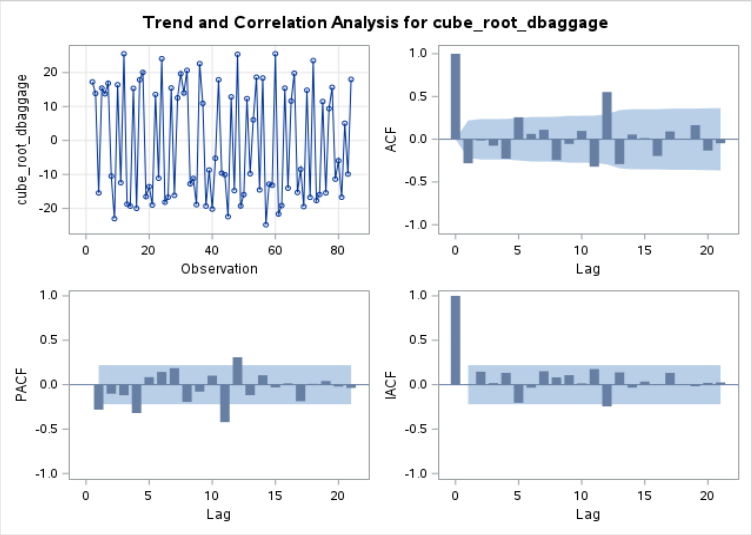
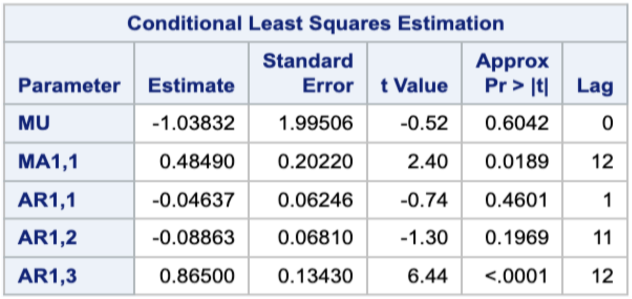


Figure 14: Trend and Correlation Analysis for P = (1 12) and Q = (5 11 12)



Conditional Least Squares Estimation					
Parameter	Estimate	Standard Error	t Value	Approx Pr > |t|	Lag
MU	-1.03832	1.99506	-0.52	0.6042	0
MA1,1	0.48490	0.20220	2.40	0.0189	12
AR1,1	-0.04637	0.06246	-0.74	0.4601	1
AR1,2	-0.08863	0.06810	-1.30	0.1969	11
AR1,3	0.86500	0.13430	6.44	<.0001	12

Figure 15: Significance of lags using P = (1 11 12) and Q = (12)

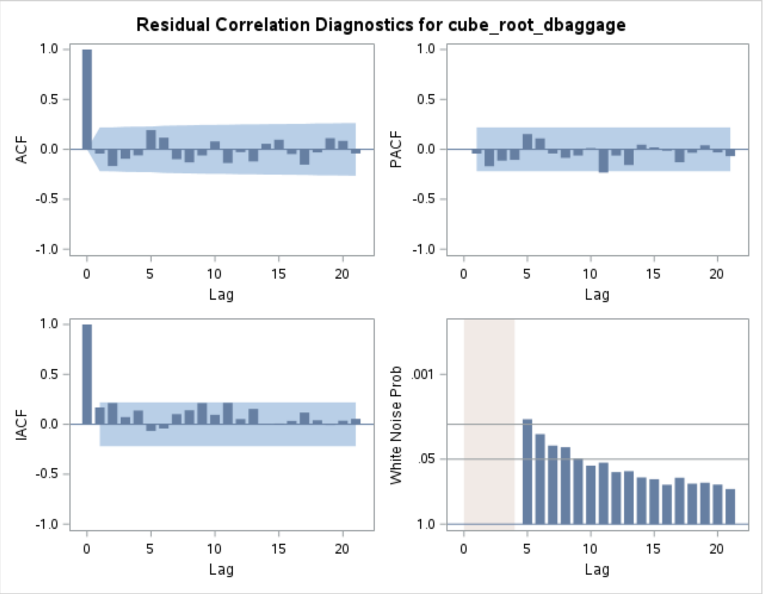


Figure 16: Residual Diagnostics for P = (11 12) and Q = (5 12)

Results

The equation for American Eagle Airlines is:

$$\sqrt[3]{z_t} + \ 0.11527\sqrt[3]{z_{t-11}} - 0.83069\sqrt[3]{z_{t-12}} = a_t - 0.23255\ a_{t-12}$$

which indicates that we took a cube root of the first difference, along with AR(11,12) and MA(12) lags.

We generated a 5-year forecast using this model in Figure 17. Seasonality is present in the forecast as a period seems to occur every twelve observations (i.e., every year) in which peaks of baggage claims can be found during the summer and winter months, due to increased travel and an influx of passengers.

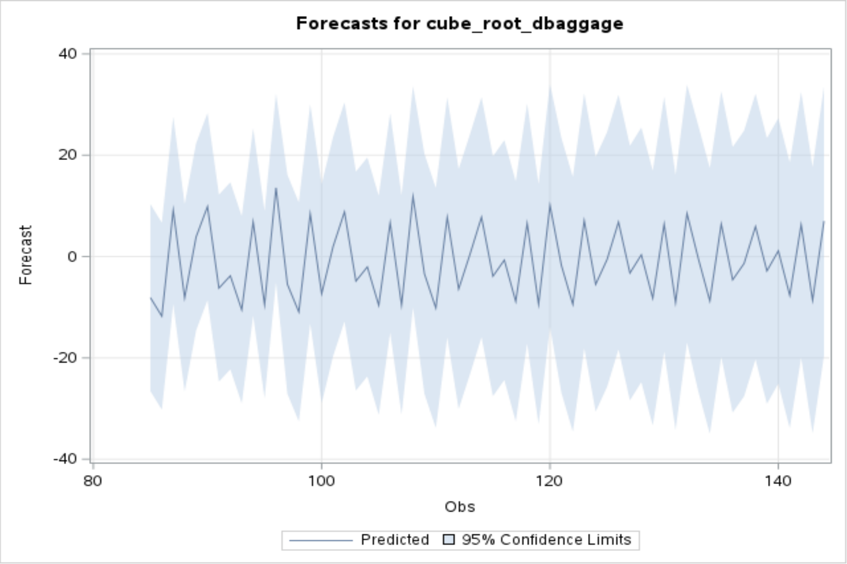


Figure 17: 5-year forecast for baggage complaints for American Eagle Airlines

The equation for United Airlines is:

$$\sqrt[3]{z_t} + \ 0.06133\sqrt[3]{z_{t-11}} - 0.93867\sqrt[3]{z_{t-12}} = a_t - 0.12464\ a_{t-5} - 0.58448\ a_{t-12}$$

which indicates that we took a cube root of the first difference along with AR(11,12) and MA(5,12) lags.

We generated a 5-year forecast using this model in Figure 18. Like American Eagle Airlines, seasonality is also present in the forecast for United Airlines as a period seems to occur every 12 observations (i.e. every year) in which peaks of baggage claims can be found during the summer and winter months, most likely due to increased travel and an influx of passengers. However, for United, the peaks are higher, due to their larger range of customers which would lead to a larger opportunity for claims.

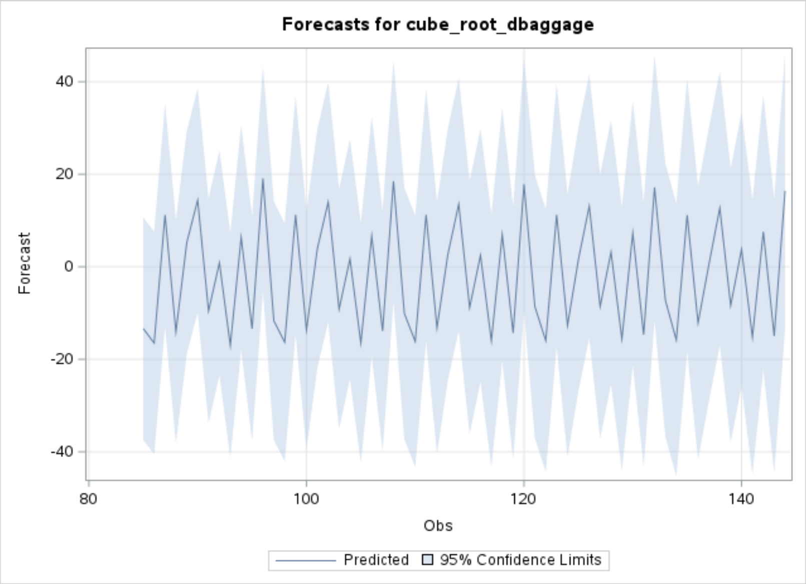


Figure 18: 5-year forecast for baggage complaints for United Airlines

Summary

In this analysis we aimed to create a model using previous data of reported lost luggage in the airline industry during the 2004-2010 period. More specifically our focus centered around two airlines, United and American Eagle. After using modeling procedures for the American Eagle Airlines data, our model displayed key peaks indicating the high volume of baggage claims during the summer and winter months. As shown in our 5-year forecast, there were multiple peaks corresponding to a surge in travel relating to more complex airline logistics. The United airlines model displayed similar characteristics depicting the seasonality of baggage claims. One distinct difference was that the United airline peaks were larger, correlating to their more extensive customer pool. While these results and forecasts are relevant to their respective airlines, we should not generalize these findings to the entire industry as fleet sizes and operational routes can play a key role in affecting baggage claims. Additionally, these forecasts do not consider confounding factors such as technological advancements that might reduce the amount of baggage claims made in upcoming years. However, this analysis shows that continued monitoring of industry trends can allow airlines to leverage data-driven insights to optimize baggage handling processes and ensure seamless travel experiences for passengers.

References

"American Eagle Airlines Fleet Details | Airfleets Aviation." Airfleets.Net, Airfleets.net, www.airfleets.net/flottecie/American%20Eagle.htm. Accessed 29 Apr. 2024.

"Baggage Mishandled by Marketing U.S. Air Carriers." Bureau of Transportation Statistics, United States Department of Transportation, 19 Apr. 2023, www.bts.gov/content/mishandled-baggage-reports-filed-passengers-largest-us-air-carriersa.

Caterinicchia, Dan. "Airline Delays in 2007 Were Second Worst Ever, With More Than 26 Percent of Flights Late." Savannah Morning News, Savannah Morning News, 6 Feb. 2008, www.savannahnow.com/story/news/2008/02/06/airline-delays-2007-were-second-worst-ever-more-26-percent-flights-late/13774463007/.

"Envoy Air." Wikipedia, Wikimedia Foundation, 26 Apr. 2024, en.wikipedia.org/wiki/Envoy_Air.

"Full Year 2022 U.S. Airline Traffic Data." Bureau of Transportation Statistics, United States Department of Transportation, 16 Mar. 2023, www.bts.gov/newsroom/full-year-2022-us-airline-traffic-data#:~:text=U.S.%20airlines%20carried%20194%20million,and%20388%20million%20in%202020.

"Hawaiian Airlines Fleet Details | Airfleets Aviation." Airfleets.Net, Airfleets.net, www.airfleets.net/flottecie/Hawaiian%20Airlines.htm. Accessed 29 Apr. 2024.

"Hawaiian Airlines." Wikipedia, Wikimedia Foundation, 27 Apr. 2024, en.wikipedia.org/wiki/Hawaiian_Airlines.

Santello, G. (2023, September 28). Airline Baggage Complaints - Time Series Dataset. Kaggle. https://www.kaggle.com/datasets/gabrielsantello/airline-baggage-complaints-time-series-dataset?select=baggagecomplaints.csv

"Star Alliance Facts and Figures" (PDF). Star Alliance. March 31, 2014. Archived from the original (PDF) on 16 Oct. 2015. Retrieved 4 Apr. 2014.

"The Boeing Logbook: 1927–1932". Boeing. Archived from the original on 7 Jan. 2015. Retrieved 3 Dec. 2014.

"United Airlines Fleet Details | Airfleets Aviation." Airfleets.Net, Airfleets.net, www.airfleets.net/flottecie/United%20Airlines.htm. Accessed 29 Apr. 2024.

Appendix

```
PROC IMPORT OUT=baggage
   DATAFILE="/home/u63069202/Time Series Analysis/Project/baggagecomplaints.csv"
   DBMS=CSV
   REPLACE;
   GETNAMES=YES;
RUN;

data baggage;
   set work.baggage;
   where airline ne 'Hawaiian';
run;

data work.baggage;
   set work.baggage;
   by airline;
   if first.airline then do;
      dbaggage = .; /* Initialize the first difference for each airline */
      prev_baggage = .; /* Initialize variable to hold previous baggage value */
   end;
if first.airline then output; /* Output only when it's the first observation for each airline */
   else do;
      dbaggage = dif(baggage); /* Calculate the first difference using the DIF function */
      cube_root_dbaggage = sign(dbaggage) * abs(dbaggage) ** (1/3); /* Take the cube root of
the first difference */
                  departed_passengers = Enplaned/(Scheduled - Canceled);  /* departed
passengers per flight */
     output;
   end;
   prev_baggage = baggage; /* Store current baggage value for the next iteration */
run;

data American United;
   set work.baggage;
   if airline = "American Eagle" then output American;
   else if airline = "United" then output United;
run;

proc sgplot data=baggage;
         series x=date y=baggage / group = airline;
run;

proc sgplot data=baggage;
         series x=date y=dbaggage / group = airline;
run;
```

```
proc sgplot data=baggage;
        series x=date y=cube_root_dbaggage / group = airline;
run;

/* CCF  */

data baggage_indicator;
   set baggage;
   if airline = 'United' then united = 1;
   else united = 0;
   if airline = 'American Eagle' then american = 1;
   else american = 0;
run;

proc arima data=baggage_indicator;
identify var=cube_root_dbaggage crosscor=(united american) nlag=21 scan;
estimate p=(11 12) q=(12) noint;
run;

/* only American  */

PROC AUTOREG DATA=american PLOTS(ONLY)= (ACF WN PACF);
                        MODEL cube_root_dbaggage = / partial NLAG=21 method=ml backstep;
RUN;

PROC ARIMA DATA=american;
        IDENTIFY VAR=cube_root_dbaggage SCAN ;
        RUN;

PROC ARIMA DATA=american;
        IDENTIFY VAR=cube_root_dbaggage nlag=21;
        ESTIMATE P=(1 12) Q=(12) PRINTALL PLOT; /* found through autoreg */
        FORECAST LEAD=60;
RUN;

PROC ARIMA DATA=american;
        IDENTIFY VAR=cube_root_dbaggage nlag=21;
        ESTIMATE P=1 Q=0 PRINTALL PLOT; /* found through scan */
        FORECAST LEAD=60;
RUN;

PROC ARIMA DATA=american;
        IDENTIFY VAR=cube_root_dbaggage nlag=21;
        ESTIMATE P=(1 11 12) Q=(12) PRINTALL PLOT; /* found through autoreg plots*/
        FORECAST LEAD=60;
RUN;
```

```
PROC ARIMA DATA=american;
        IDENTIFY VAR=cube_root_dbaggage nlag=21;
        ESTIMATE P=(11 12) Q=(12) PRINTALL PLOT; /* found through autoreg sig. */
        FORECAST LEAD=60;
RUN;

/* only United  */

PROC AUTOREG DATA=united PLOTS(ONLY)= (ACF WN PACF);
                        MODEL cube_root_dbaggage = / partial NLAG=21 method=ml backstep;
RUN;

PROC ARIMA DATA=united;
        IDENTIFY VAR=cube_root_dbaggage SCAN ;
        RUN;

PROC ARIMA DATA=united;
        IDENTIFY VAR=cube_root_dbaggage nlag=21;
        ESTIMATE P=(1 12) Q=(5 11 12) PRINTALL PLOT; /* found through autoreg */
        FORECAST LEAD=60;
RUN;

PROC ARIMA DATA=united;
        IDENTIFY VAR=cube_root_dbaggage nlag=21;
        ESTIMATE P=2 Q=2 PRINTALL PLOT; /* found through scan */
        FORECAST LEAD=60;
RUN;

PROC ARIMA DATA=united;
        IDENTIFY VAR=cube_root_dbaggage nlag=21;
        ESTIMATE P=(1 11 12) Q=(12) PRINTALL PLOT; /* found through autoreg plots*/
        FORECAST LEAD=60;
RUN;

PROC ARIMA DATA=united;
        IDENTIFY VAR=cube_root_dbaggage nlag=21;
        ESTIMATE P=(11 12) Q=(12) PRINTALL PLOT; /* found through autoreg sig. */
        FORECAST LEAD=60;
RUN;

PROC ARIMA DATA=united;
        IDENTIFY VAR=cube_root_dbaggage nlag=21;
        ESTIMATE P=(11 12) Q=(5 12) PRINTALL PLOT; /* found through autoreg AIC */
        FORECAST LEAD=60;
RUN;
```