
AMERICAN AIRLINES
BAG CHECK PREDICTION
Alyssa Juarez

THE GOAL

The Goal of this project is to create an algorithm that can accurately predict the number of bags that will be on each flight from American Airlines.


PREPROCESSING

BASIC DATA CLEANING

01

REMOVE NULLS & DUPLICATES

Remove or replace any empty values & remove duplicate rows that occur in the set.

02

TRIM FEATURES

Trim Features down to only those that are relevant for our purposes.

03

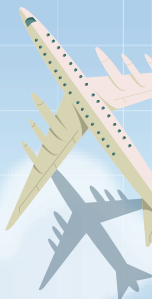
DATA GROUPING

Find the best way to group the data so that we can see what we can call normal and how each variable is effected

04

REMOVE OUTLIERS

Remove any anomalies.



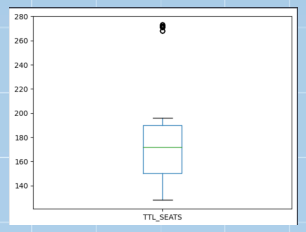
MY DATASET

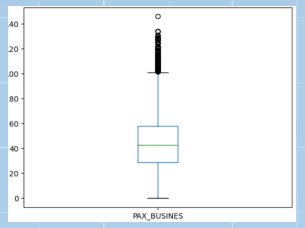
The original dataset contained 74 variables, and 258,920 instances. Here are are the features I found relevant

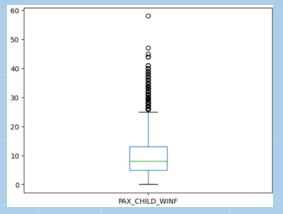
TTL_SEATS	Total number of seats available for the given aircraft type
TTL_PAX	Total number of passengers that traveled in the given flight
PAX_BUISNESS	Number of passengers travelling for business purposes booked through -third party travel agencies
PAX_LEISR	Number of passengers traveling for leisure - booked by third party, e.g., expedia
PAX_CHILD_WINF	Number of passengers travelling with child or infants
PAX_ASSIST	Number of passengers needing special assistance, e.g., wheel chair, etc.
PAX_ANIMAL	Number of passengers travelling with an animal
CABIN_FIRST	Number of passengers seating in First Class cabin
CABIN_BUISNESS	Number of passengers seating in First Class cabin
CABIN)PREM_ECON	Number of passengers seating in Premium Economy cabin
CABIN_MAIN	Number of passengers seating in Main or economy cabin
FLIGHT_ARVL_DT	Arrival date for flight departing from CLT
DEST_CITY_NM	Destination city name
TTL_BAGS	Total number of checked bags in the flight

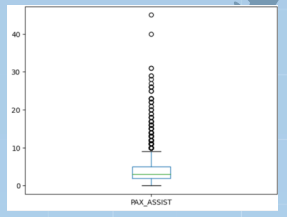
OUTLIERS

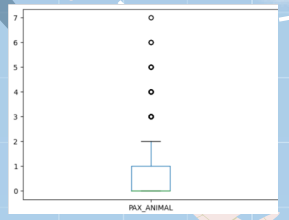
My dataset is very complicated with outliers, I will need to break apart the dataset into groups in order to get a more reliable read for what is normal. As removing outliers outright results in the removal of too many of my features.

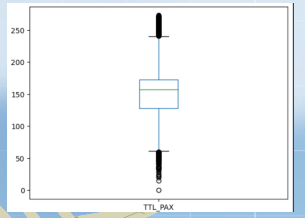


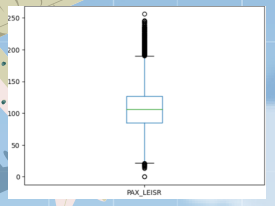


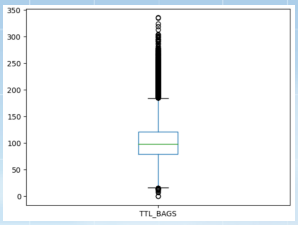


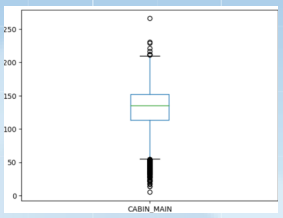


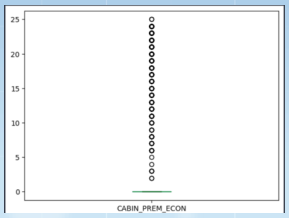


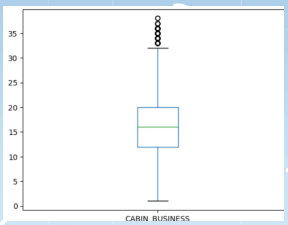


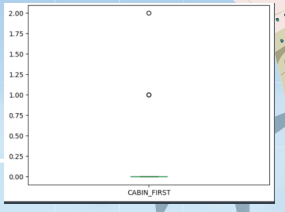






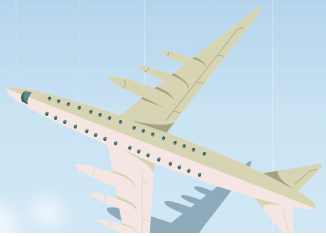






DATA LOSS PER CLEANING STEP

Original size	258920 rows
After dropping null values	255982 rows
After dropping duplicates	33335 rows
After dropping outliers	11028 rows



FEATURE CREATION

01 SEASON

The season in which the flight occurs. Colder weather may encourage more baggage.

02 PROX_XMAS

Proximity to Christmas

03 PROX_THXG

Proximity to Thanksgiving

04 BACK_TO_SCHOOL

Proximity to the first day back to school

05 END_OF_SCHOOL

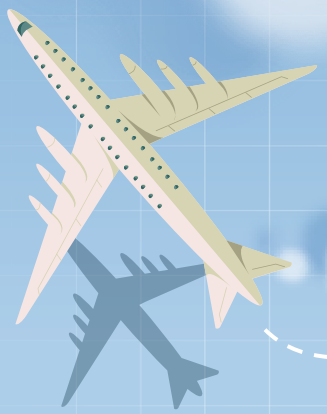
Proximity to the beginning of summer break

06 TEMP

The temperature of the travel destination

FEATURE CREATION

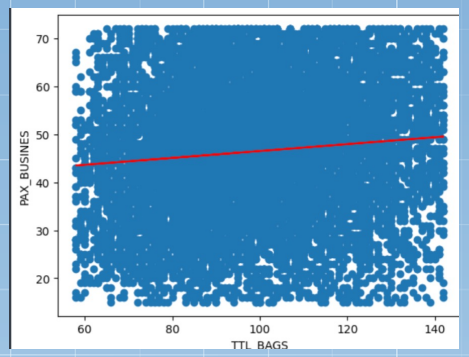
After some Analysis, I realized I was unable to do several of the features I would like to with this dataset given that the data only covers May to September.

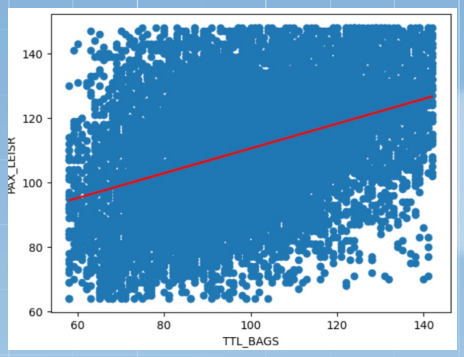


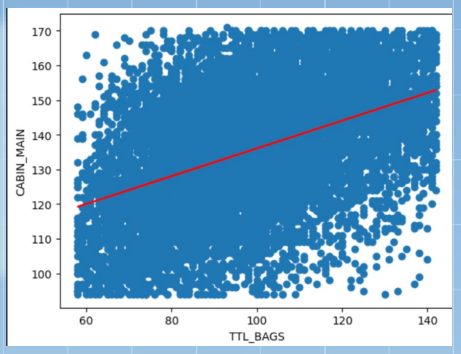
```python
# END OF SCHOOL

df_cat['EOS1'] = '2022-05-08'
df_cat['EOS2'] = pd.to_datetime(df_cat['EOS1'])
df_cat['FAD'] = pd.to_datetime(df_cat['FLIGHT_ARVL_DT'])
df_cat.drop(['EOS1'], axis = 1)
df_cat['EOS'] = (df_cat['FAD'] - df_cat['EOS2']) / np.timedelta64(1, 'D
dfcat = df_cat.drop(['EOS1', 'FAD', 'EOS2'], axis = 1)
dfcat
```

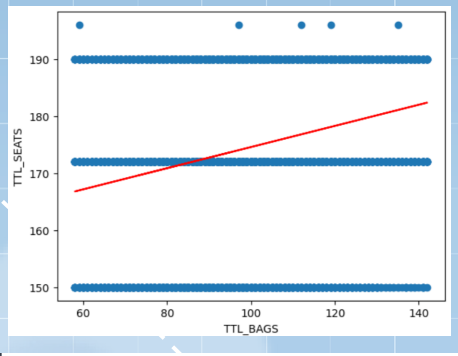
```python
# BEGINNING OD SCHOOL
dfcat['BOS1'] = '2022-08-22'
dfcat['BOS2'] = pd.to_datetime(dfcat['BOS1'])
dfcat['FAD'] = pd.to_datetime(dfcat['FLIGHT_ARVL_DT'])
dfcat.drop(['BOS1'], axis = 1)
dfcat['BOS'] = (dfcat['FAD'] - dfcat['BOS2']) / np.timedelta64(1, 'D')
dfcat2 = dfcat.drop(['BOS1', 'FAD', 'BOS2'], axis = 1)
dfcat2
```

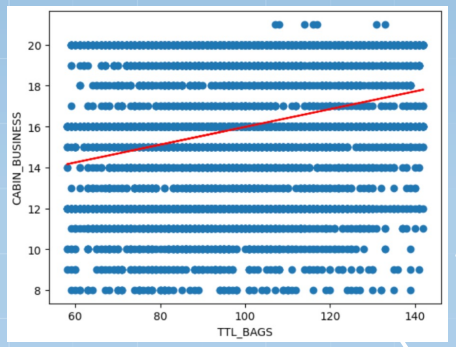

DATA ANALYSIS

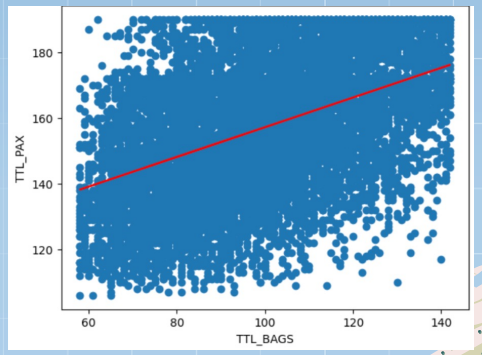

PAX_BUSINES
TTL_BAGS

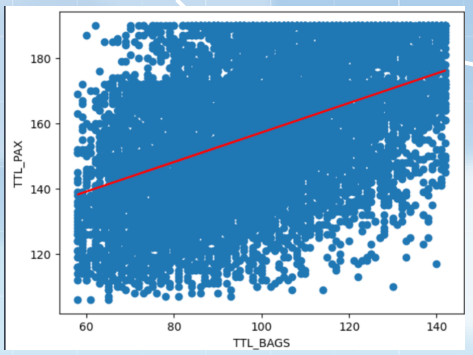

PAX_LEISR
TTL_BAGS

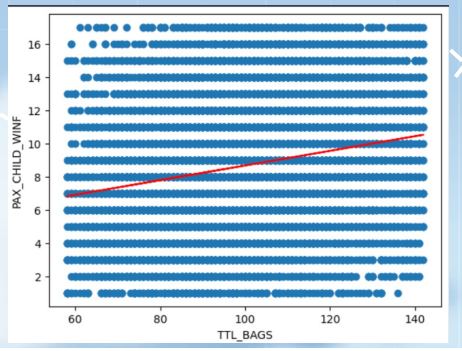

CABIN_MAIN
TTL_BAGS

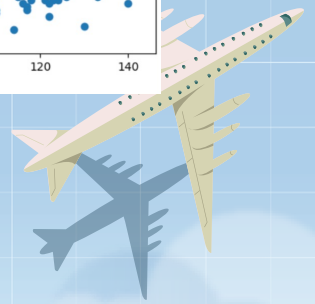

TTL_SEATS
TTL_BAGS

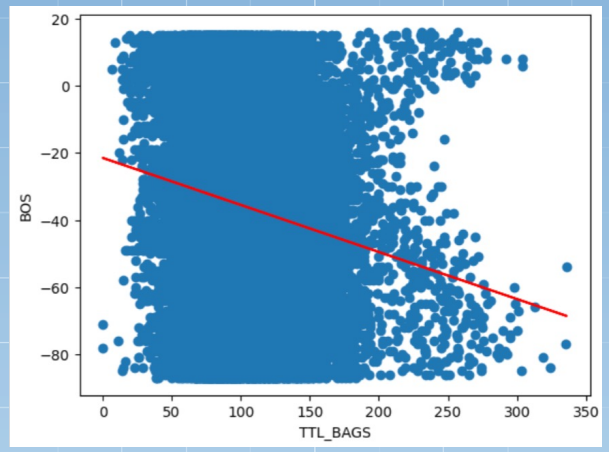

CABIN_BUSINESS
TTL_BAGS

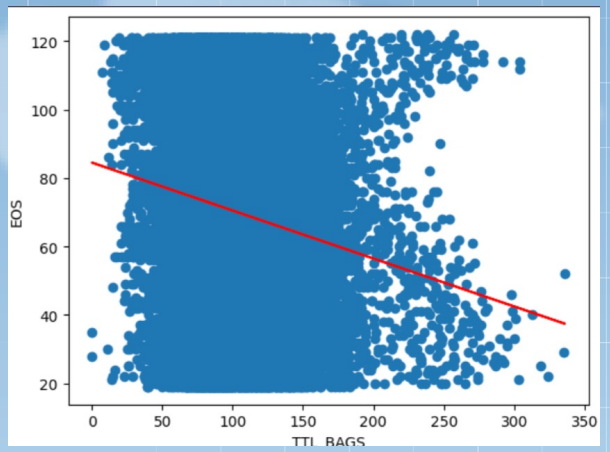

TTL_PAX
TTL_BAGS

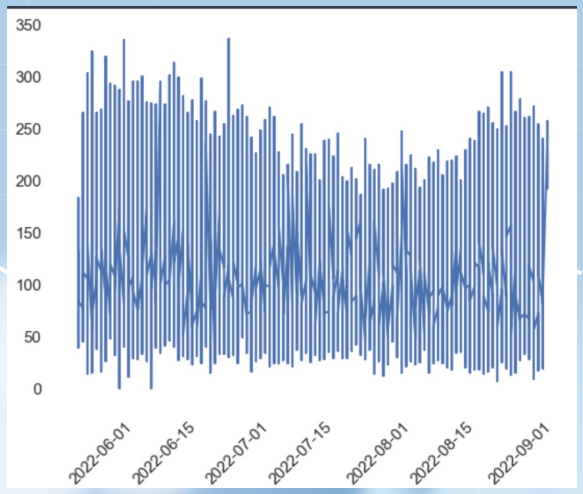

TTL_PAX
TTL_BAGS

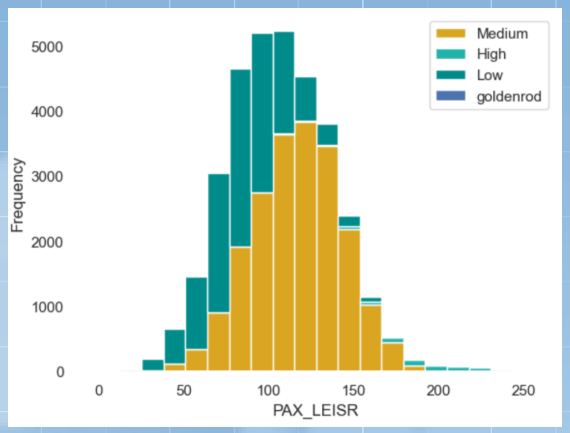

PAX_CHILD_WINF
TTL_BAGS

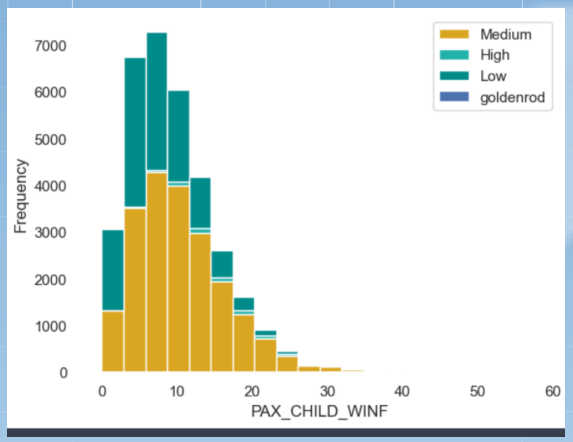


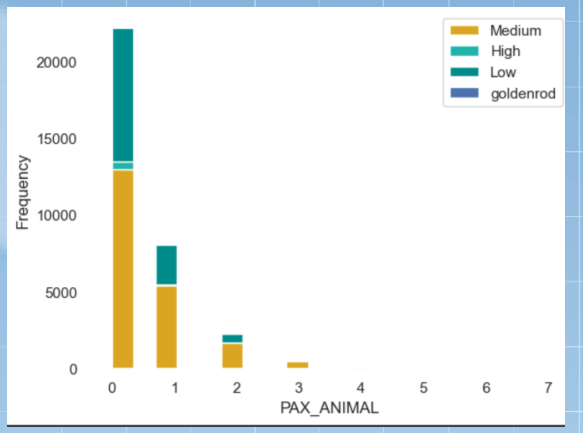

BOS
TTL_BAGS
20
0
−20
−40
−60
−80
0
50
100
150
200
250
300
350

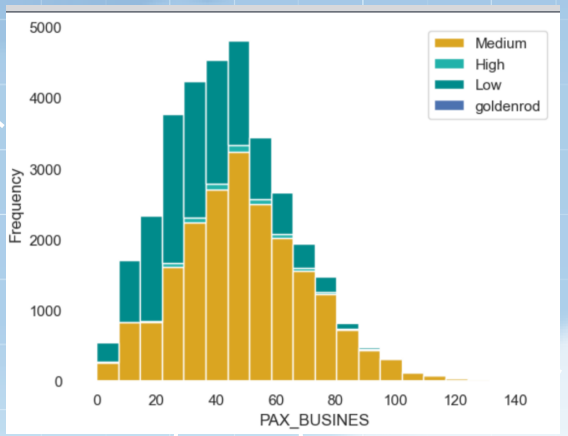

EOS
TTL_BAGS
120
100
80
60
40
20
0
50
100
150
200
250
300
350

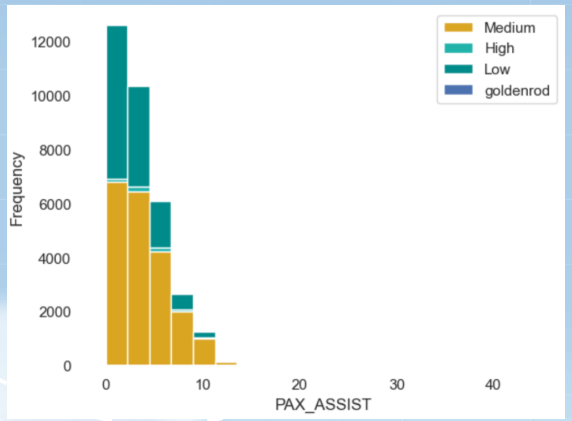

350
300
250
200
150
100
50
0
2022-06-01
2022-06-15
2022-07-01
2022-07-15
2022-08-01
2022-08-15
2022-09-01

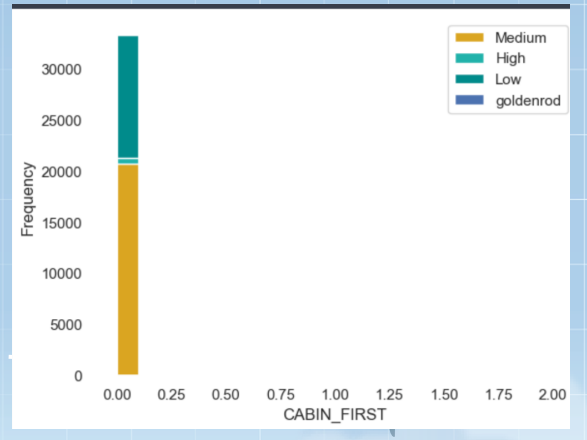

Medium
High
Low
goldenrod
Frequency
PAX_LEISR

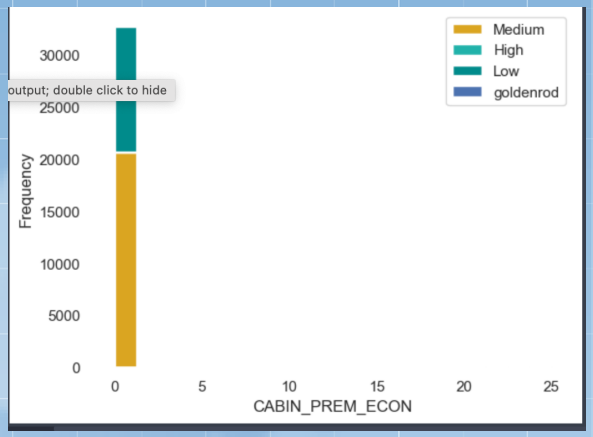

Medium
High
Low
goldenrod
Frequency
PAX_CHILD_WINF

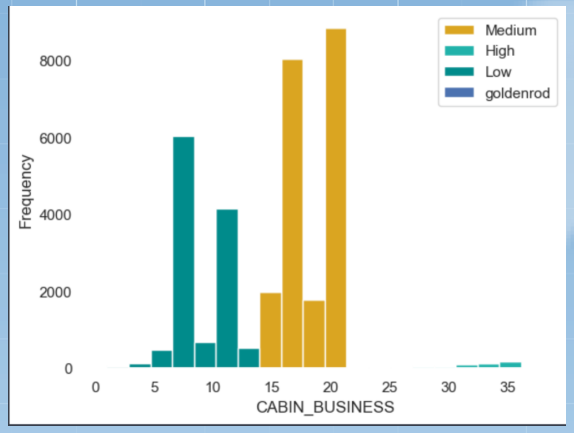

Medium
High
Low
goldenrod
Frequency
PAX_ANIMAL

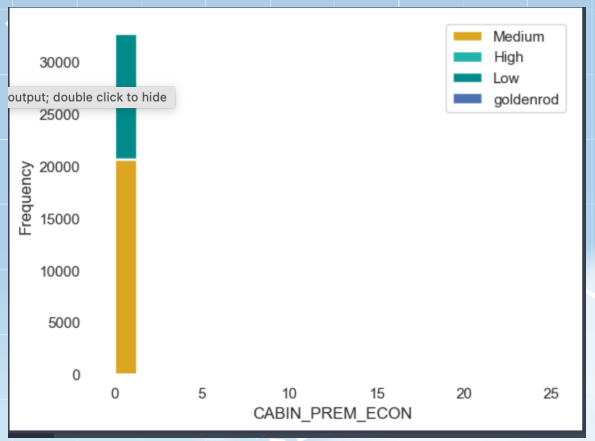

Medium
High
Low
goldenrod
Frequency
PAX_BUSINES

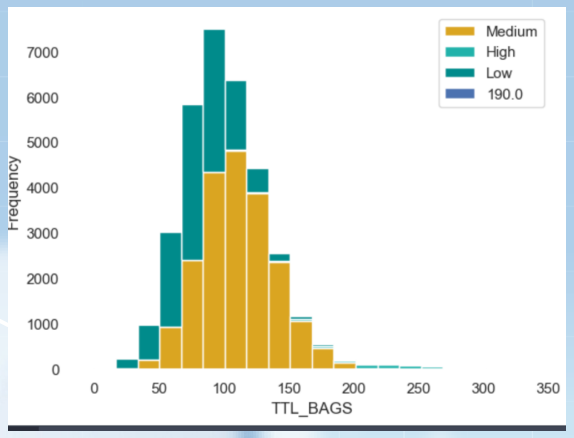

Medium
High
Low
goldenrod
Frequency
PAX_ASSIST


Medium
High
Low
goldenrod
Frequency
CABIN_FIRST


Medium
High
Low
goldenrod
output; double click to hide
Frequency
30000
25000
20000
15000
10000
5000
0
0
5
10
15
20
25
CABIN_PREM_ECON


Medium
High
Low
goldenrod
Frequency
8000
6000
4000
2000
0
0
5
10
15
20
25
30
35
CABIN_BUSINESS


Medium
High
Low
goldenrod
output; double click to hide
Frequency
30000
25000
20000
15000
10000
5000
0
0
5
10
15
20
25
CABIN_PREM_ECON


Medium
High
Low
190.0
Frequency
7000
6000
5000
4000
3000
2000
1000
0
0
50
100
150
200
250
300
350
TTL_BAGS

CHECKING FOR MULTICOLLINEARITY

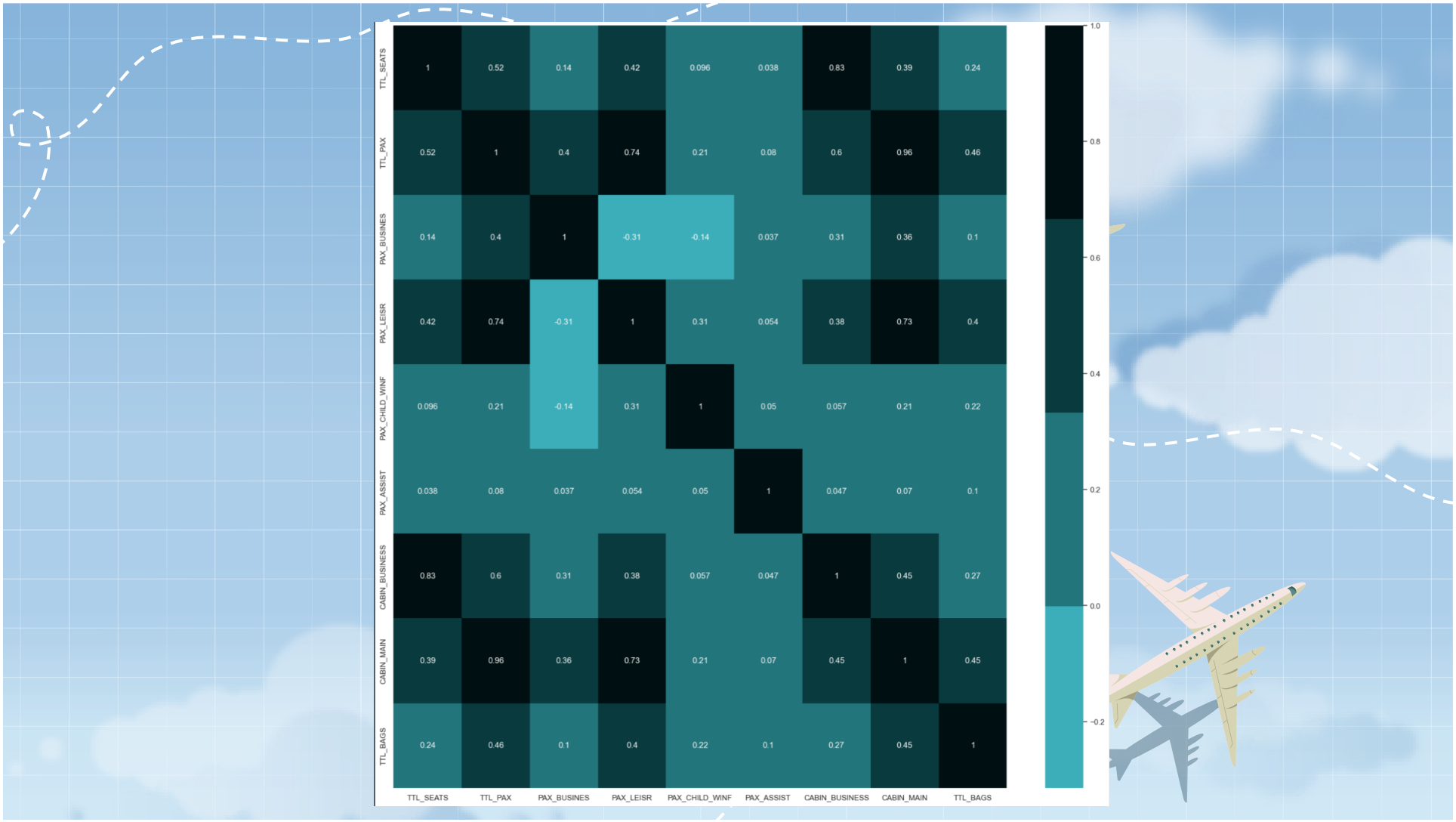
Variance Inflation Factor (VIF) Results:

	Variable	VIF
0	TTL_SEATS	60.278522
1	TTL_PAX	5618.070894
2	PAX_BUSINES	454.811530
3	PAX_LEISR	2347.425826
4	PAX_CHILD_WINF	4.987747
5	PAX_ASSIST	3.519100
6	PAX_ANIMAL	1.524878
7	CABIN_FIRST	1.004760
8	CABIN_BUSINESS	44.406449
9	CABIN_PREM_ECON	1.803345
10	CABIN_MAIN	805.373017
11	TTL_BAGS	25.618489

Correlation Coefficients

	TTL_SEATS	TTL_PAX	PAX_BUSINES	PAX_LEISR	PAX_CHILD_WINF
TTL_SEATS	1.000000	0.517799	0.144910	0.417251	0.095725
TTL_PAX	0.517799	1.000000	0.396037	0.743222	0.205118
PAX_BUSINES	0.144910	0.396037	1.000000	-0.314204	-0.138541
PAX_LEISR	0.417251	0.743222	-0.314204	1.000000	0.309929
PAX_CHILD_WINF	0.095725	0.205118	-0.138541	0.309929	1.000000
PAX_ASSIST	0.037917	0.079747	0.037475	0.053932	0.050298
CABIN_BUSINESS	0.833595	0.601406	0.314189	0.384609	0.056823
CABIN_MAIN	0.386758	0.956760	0.355939	0.730735	0.208656
TTL_BAGS	0.244905	0.459196	0.101788	0.398418	0.215129

	PAX_ASSIST	CABIN_BUSINESS	CABIN_MAIN	TTL_BAGS
TTL_SEATS	0.037917	0.833595	0.386758	0.244905
TTL_PAX	0.079747	0.601406	0.956760	0.459196
PAX_BUSINES	0.037475	0.314189	0.355939	0.101788
PAX_LEISR	0.053932	0.384609	0.730735	0.398418
PAX_CHILD_WINF	0.050298	0.056823	0.208656	0.215129
PAX_ASSIST	1.000000	0.047192	0.070195	0.103151
CABIN_BUSINESS	0.047192	1.000000	0.453829	0.267997
CABIN_MAIN	0.070195	0.453829	1.000000	0.446952
TTL_BAGS	0.103151	0.267997	0.446952	1.000000


TTL_SEATS
TTL_PAX
PAX_BUSINES
PAX_LEISR
PAX_CHILD_WINF
PAX_ASSIST
CABIN_BUSINESS
CABIN_MAIN
TTL_BAGS
1 0.52 0.14 0.42 0.096 0.038 0.83 0.39 0.24
0.52 1 0.4 0.74 0.21 0.08 0.6 0.96 0.46
0.14 0.4 1 -0.31 -0.14 0.037 0.31 0.36 0.1
0.42 0.74 -0.31 1 0.31 0.054 0.38 0.73 0.4
0.096 0.21 -0.14 0.31 1 0.05 0.057 0.21 0.22
0.038 0.08 0.037 0.054 0.05 1 0.047 0.07 0.1
0.83 0.6 0.31 0.38 0.057 0.047 1 0.45 0.27
0.39 0.96 0.36 0.73 0.21 0.07 0.45 1 0.45
0.24 0.46 0.1 0.4 0.22 0.1 0.27 0.45 1
TTL_SEATS
TTL_PAX
PAX_BUSINES
PAX_LEISR
PAX_CHILD_WINF
PAX_ASSIST
CABIN_BUSINESS
CABIN_MAIN
TTL_BAGS
1.0
0.8
0.6
0.4
0.2
0.0
−0.2


THE NEXT STEPS

THE PLAN

Over the break I plan to finish my data analysis, I suspect I will need to create groups within the data set in order to get a more reliable view into how each variable effects the total bag count.

Next semester I would like to begin with building my model and testing its accuracy. I also would like to reach out to American Airlines to test the model on multiple Airports data.