Project 3 - Queueing Systems

Celine Chiou(Yun-Chyau, Chiou) 300230747, Xinyu Xie 0300239933, Deonne Millaire 300260913

November 19, 2021

Introduction

The Borealian Aeronautic Security Agency (BASA) runs pre-board screening of passengers and crew for all flights departing the nation's airfields. The four major airfields are: Auckland, Chebucto, Saint-François and Queenston. In the following project, we are going to predict the wait time at each airfields.

The screening process (PBS) is structurally similar at each airfield:

1. Passengers arrive at the beginning of the main queue
2. Boarding passes may or may not be scanned at S_1
3. Passengers enter the main queue
4. Boarding passes are scanned at S_2
5. Passengers are directed to a server entry position
6. Passengers and carry-on luggage are screened by a server

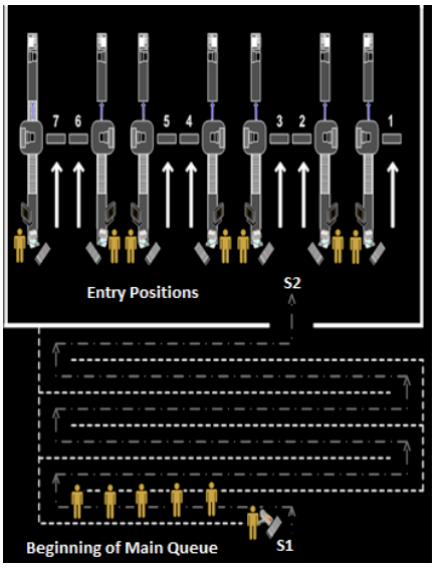


Figure 1: Screening Process

Initial Exploratory data analysis

There are four datasets we can choose to conduct the analysis. We will analyze and compare each one of these in this report.

Data_F contains information for a given flight and for example the flight capacity and actual number of passengers for each flight that has already taken off. The content of Data_P and BASA_AUC are similar, they both contain information about each passenger. After we joined these two tables by a common variable called Pass_ID (which represents the passenger ID), and check whether date/time through the S2 checkpoint is the same, we conclude that the BASA_AUC and Data_p dataset contain information regarding the same passengers, however BASA_AUC has 30 more observations compared with Data_p. Furthermore, we hypothesize that the BASA_AUC data set contains the raw data. In the BASA data set, it contained many missing values in the C_Start and the C_avg variables. Whenever these values were NA, the data_P data set actually had data in these positions. Therefore we believe at some point someone took the original BASA data and somehow imputed missing values for the missing C_Start values. As we do not have the original data sources for these data sets, we have decided to choose to work with the Dat_P data set going forward.

After conducting missing value analysis, we found that whenever C_start is NA, Wait_Time and C_avg is NA. However, there is no NA for C0. So we guess that C_Start is the number of active servers when a passenger enters the main queue, and C0 is the number of active servers when the passenger reaches the end of main queue (S2). Also we have verified that C_avg is indeed the average of C_0 and C_start. We simply choose the C_avg for our following calculation.

The last dataset contains information for 4 different airfields(Auckland, Chebucto, Saint-Francois, Queenston). According to the distribution plot, we noticed that 99.6% of observations are related to Saint-Francois; there are only 4, 1, 7 observations for Auckland, Chebucto, Queenston respectively. Among these 3214 observations for the Saint Francois airfield in January 2030, there are 41.82% missing values for the variable C_avg.

Task 1

Data dictionary

Basa_AUC_2028_912/years20262030

Table 1: Basa_AUC_2028_912/years20262030

Variable Name	Variable Discription	Variable Type
X	Passenger number	0
Airfield	Which of the 4 airfields the observation is about,	1
S2	The wait time between s1 and s2	0
Wait_Time	Date/Time that a passenger entered the S2 checkpoint	0
C_Start	Number of servers (when the passenger arrive S1)	0
C0	Number of servers (when the passenger arrive S2	0
C_avg	Average of Cstart & C0	0
Sch_Departure	The date/time a flight was supposed to depart	0
Act_Departure	The actual date/time a flight departed	0
BFO_Dest_City	The city code (3 letter country code and 3 letter city code)	1
BFO_Destination_Country_Code	The country code, one of 6 possible countries	1
order	Number of order	0
Pass_ID	Passengers' ID	0
Departure_Date	Date at which a flight/passenger departed	0
Departure_Time	Time at which a flight/passenger departed	0

Variable Name	Variable Discription	Variable Type
Time_of_Day	Time of day represented by 4 factors	1
Period_of_Week	Period of week represented by 2 factors	1
Day_of_Week	Specific day of week as factors	1
Month	Month of the year	1
Season	One of the seasons represented as factors	1
Year	Year of the data observation, ranges from 2026-2030	1

dat_F_sub

Table 2: dat_F_sub

Variable Name	Variable Discription	Variable Type
X	Passenger number	0
Airfield	Which of the 4 airfields the observation is about,	1
Flight_ID	ID of the flight	0
Sch_Departure	The date/time a flight was supposed to depart	0
Act_Departure	The actual date/time a flight departed	0
Time_of_Day	Time of day represented by 4 factors	1
Period_of_Week	Period of week represented by 2 factors	1
Day_of_Week	Specific day of week as factors	1
Month	Month of the year	1
Season	One of the seasons represented as factors	1
Year	Year of the data observation, ranges from 2026-2030	0
tot_pass	Total number of passengers	0
N	The actual number of passengers	0
min	The minimum wait time for the flight	0
mean	The mean wait time for the flight	0
median	The median wait time for the flight	0
max	The maximum wait time for the flight	0
mean_WTL	Mean time of wait time luggage	0
mean_City_Flag	mean of city flag	0
mode_BFO_Dest_City	The city code (3 letter country code and 3 letter city code)	0
sum_city_mode	Sum of city mode	0
N_of_Dest_City	Number of destination city	0
mode_BFO_Dest_Country_Code	The destination city code	1
sum_country_mode	Sum of country mode	0
N_of_Dest_Country	Number of destiantion country	0
Delay_in_Seconds	Flight delay (or early arrival) upon destination in seconds	0

data_P_sub

Table 3: dat_P_sub

Variable Name	Variable Discription	Variable Type
Pass_ID	Passenger number	0
valid_P_ID	Binary indicator if the passenger ID is valid	0
Airfield	Which of the 4 airfields the observation is about	1
S2	Date/Time that a passenger entered the S2 checkpoint	0
Wait_Time	The wait time between s1 and s2	0
C_Start	Number of servers (when the passenger arrive S1)	0

Variable Name	Variable Discription	Variable Type
C0	Number of servers (when the passenger arrive S2)	0
C_avg	Average of Cstart & C0	0
Sch_Departure	The date/time a flight was supposed to depart	0
Act_Departure	The actual date/time a flight departed	0
BFO_Dest_City	The city code (3 letter country code and 3 letter city code)	1
BFO_Destination_Country_Code	The country code, one of 6 possible countries	1
order	Number of order	0
Departure_Date	Date at which a flight/passenger departed	1
Time_of_Day	Time of day represented by 4 factors	1
Period_of_Week	Period of week represented by 2 factors	1
Day_of_Week	Specific day of week as factors	1
Month	Month of the year	1
Season	One of the seasons represented as factors	1
Year	Year of the data observation, ranges from 2026-2030	1
WT_flag	abc	0
S2_Sch_Flag	abc	0
S2_Act_Flag	abc	0
Sch_Act_Flag	abc	0
Flight_ID	ID of the flight	0
Delay_in_Seconds	Flight delay (or early arrival) upon destination in seconds	0

Task 2

Data Visualization

We expected that the passenger volume would have some influence on the screening precess(PBS) wait time.

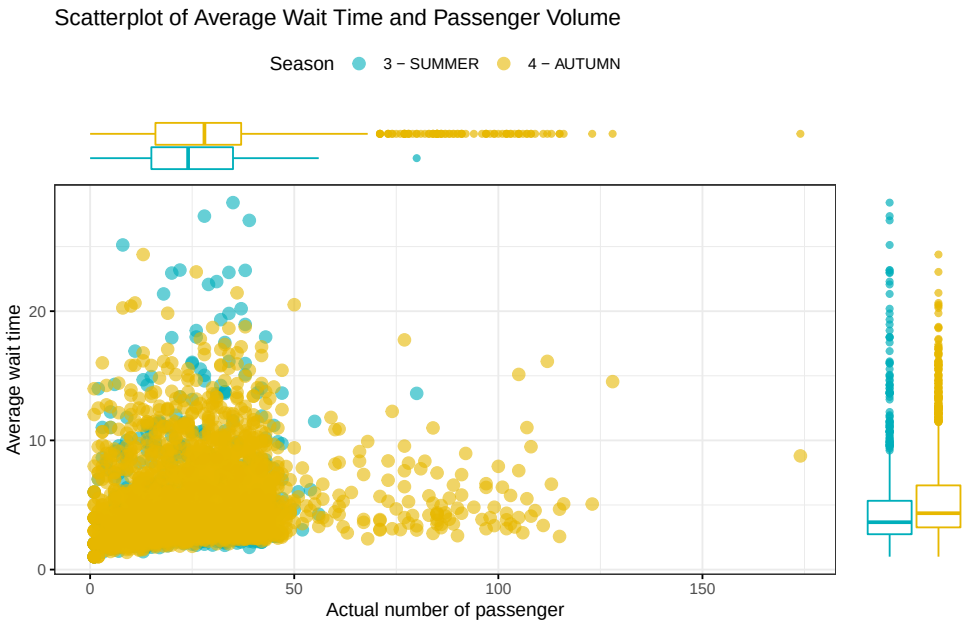


Figure 2: Average wait time verses passenger volume(period of week)

Figure 2 We try to figure out the relationship between the average wait time and passenger capacity by using scatterplot. It doesn't show a clear positive or negative relationship between these two but we can still draw the following conclusions:

- The median numbers of passengers for summer and autumn are quite similar based on the boxplots above, but the number of passengers in autumn has a wider range and has more outliers. The number of passengers in summer is basically between 0 and 50, while in winter there are a large number of observations between 50 and 100.
- The distributions for the average wait time itself are also similar for summer and autumn, the medians of wait time are approximately 5 minutes.

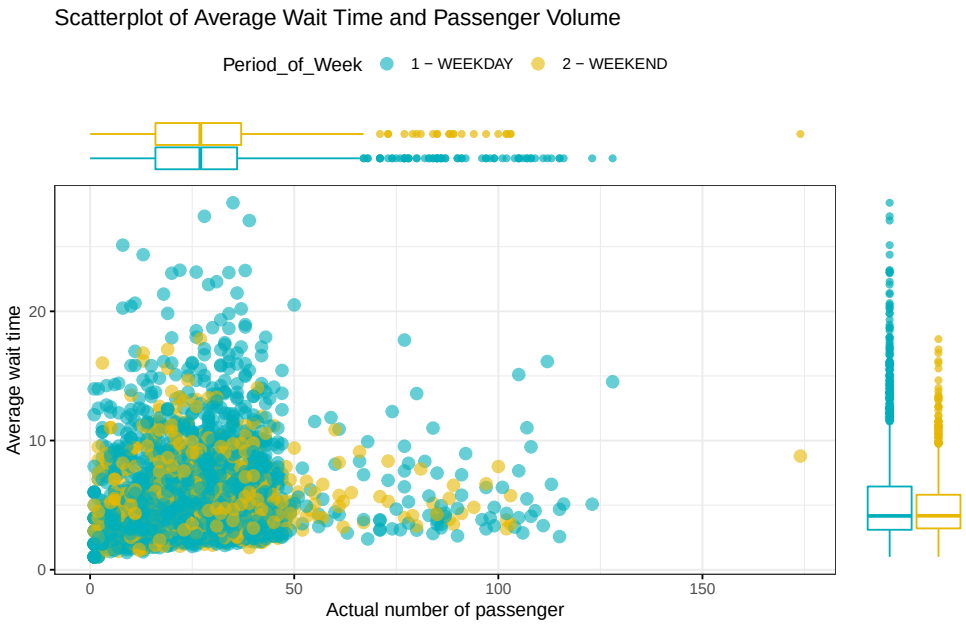


Figure 3: Average wait time verses passenger volume(Season)

Figure 3 Similar to the plot above, there is no clear pattern between the wait time and passengers.

- Median numbers of passenger are around 25 for both weekday and weekend.
- The distributions for the average wait time itself are also similar for summer and autumn, the medians of wait time are approximately 5 minutes.

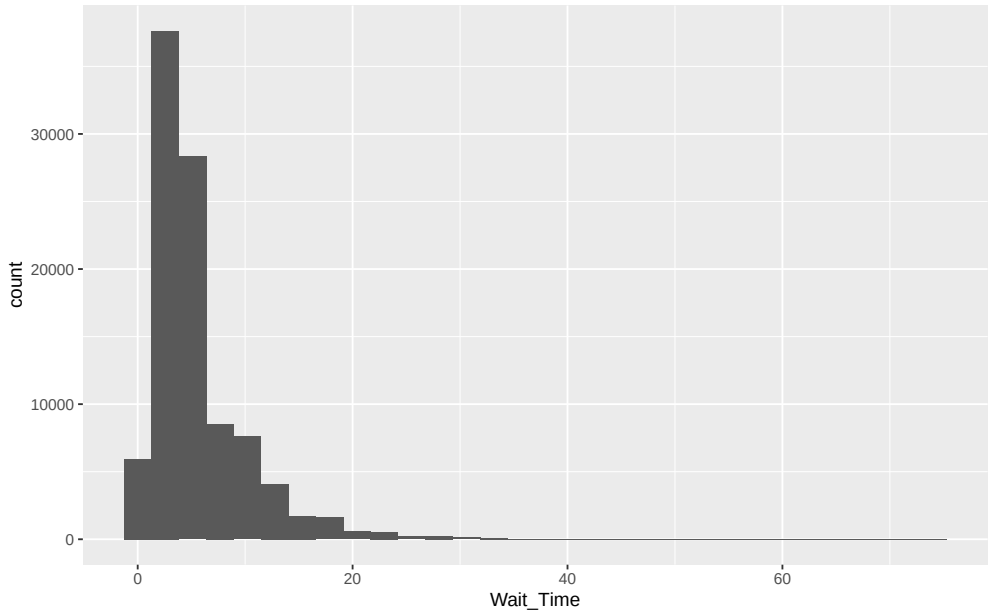


Figure 4: wait time distribution

Figure 4 According to the wait time distribution of the Basa dataset, we could conclude that almost 90% of the passengers wait within 20 minutes. And most of the passengers wait for 4 and 6 minutes between S_1 and S_2. Furthermore, except for the wait time group in 4 and 6 minutes, the rest of the wait time groups contain less than 10,000 passengers.

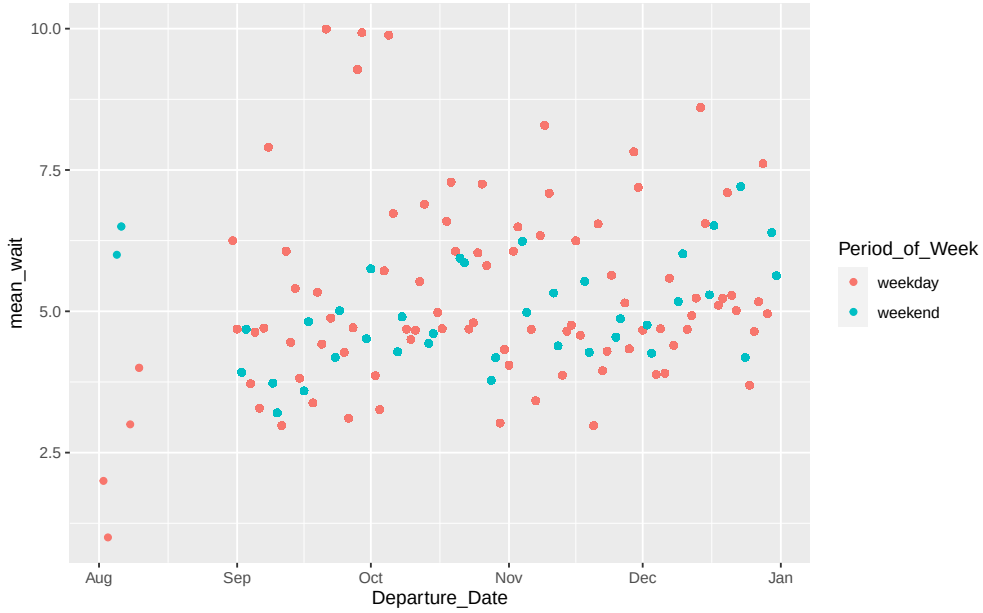


Figure 5: Departure wait verses mean wait time(Period of week)

Figure 5 As we can see from the above figure, the wait time on weekends is more stable than the wait time during the week, almost all of which fall within 3-7 minutes. Also, weekday's wait time vary from 3 minutes to 10 minutes.

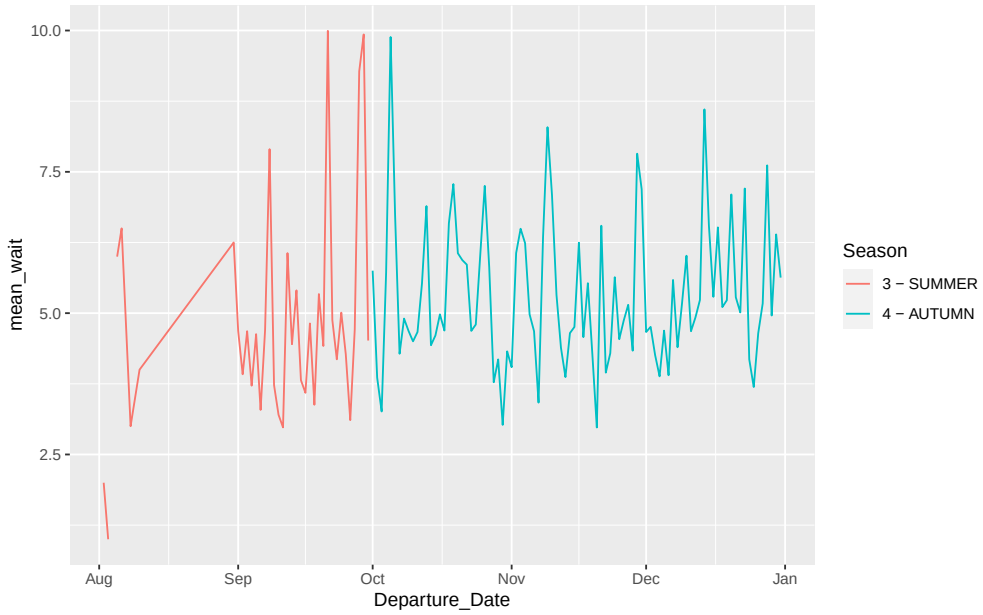


Figure 6: Departure wait verses mean wait time(Season)

Figure 6 On average, the wait time in autumn is longer than that in summer. The longest average wait time is from September to October, and the wait time after October is almost always under 7.5 minutes.

Justification of Cluster Choice

We separate our observations into different clusters based on the value of season, period of week and time of day. We assume each cluster has similar properties regarding the arrival rate, service rate, the number of servers, ect. For example, the arrival distribution for midnight and morning are significantly different from that for afternoon given the fact that people tend to purchase the flight ticket in afternoon or after 10 am, passengers tend not to arrive at their destination too early in the morning or too late in the night without consideration of budget. Moreover, the traveling population is different for weekdays and weekends. People at work tend to choose airplanes for their business travel since airplanes are a very efficient way of transportation. However, on weekends, most of the people who choose to travel by plane are because of family trips or students returning home on weekends.

Task 3

Queueing Model Qnalysis

Model

For the model, we decided to split our data into the following clusters. We first split the data in dat_P by season, which we have changed into Fall and Winter. From then we further split these two data sets into weekends and weekdays. Within these splits we then do one final split based on the time of day, which was already contained in the data set, which is morning, afternoon, evening, and night time (6 hour windows).

With these clusters, we are able to find the total number of hours, number of arrivals, arrival rate, average number of servers, the average wait time, and the service rate for each time slot of a given cluster. Below is a summary of these values of each cluster.

Imputation of Missing Wait Times

There are 16871 missing values for the variable Wait_Time in the dataset data_P. We may lose a lot of useful information by directly dropping them, we try to impute these missing values using the following method. We assume that the waiting time for a passenger will be similar to those who have a similar S2 check-in time. For example, a passenger who has checked in at 10:30am will likely share a similar wait-time as those who have checked in at 10:29am or 10:31am, as they will be close together in the queue. We impute the missing waiting time values for a given passenger by checking a plus/minus five minute window of the S2 check-in time, and taking the average wait-time from this group to be the imputed value. If there is no passenger who checked in within this five minute window, then we decided to drop the observation. We felt that the imputation was simple yet intuitive, and resulted in us not having to drop too many observations. After we impute for Wait_Time, the number of missing values decreases to 2242 from 16871.

Fall

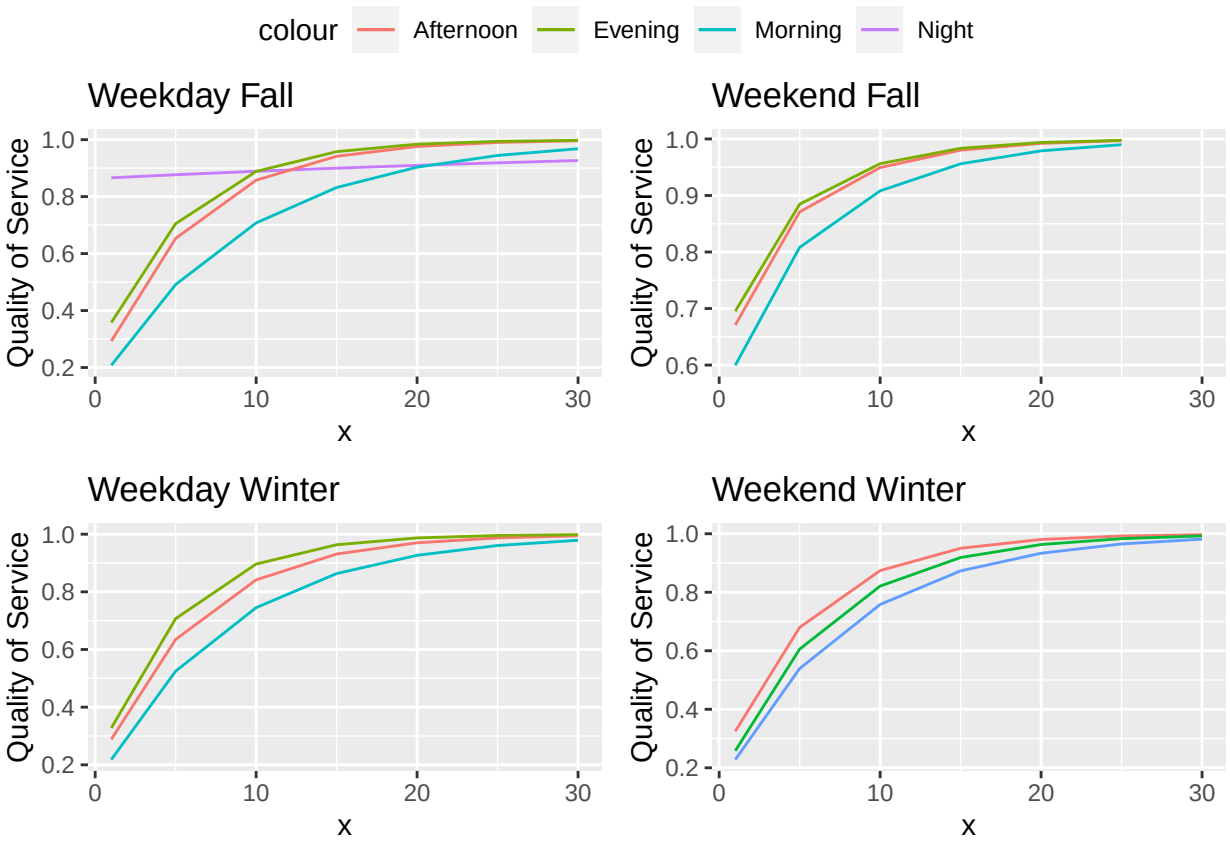
Day_Of_Week	times	num_hours	num_arrivals	arrival_rate	avg_servers	wait_time	service_rate
Weekend	Night	156	1	0.0001068	1.000000	NaN	NaN
Weekend	Morning	156	7026	0.7506410	1.460698	5.678295	0.8978722
Weekend	Afternoon	156	9181	0.9808761	1.145445	4.483559	1.1681557
Weekend	Evening	156	7490	0.8002137	1.273431	4.143604	0.9944179
Weekday	Night	390	77	0.0032906	1.123377	6.583333	0.0240628
Weekday	Morning	390	19949	0.8525214	1.525862	7.991127	0.9632725
Weekday	Afternoon	390	22706	0.9703419	1.149058	4.743870	1.1484486
Weekday	Evening	390	16126	0.6891453	1.212111	4.010192	0.8836262

Winter

Day_Of_Week	times	num_hours	num_arrivals	arrival_rate	avg_servers	wait_time	service_rate
Weekend	Morning	60	3357	0.9325000	1.479893	6.808773	1.061519
Weekend	Afternoon	60	2938	0.8161111	1.151464	4.362557	1.002682
Weekend	Evening	60	3772	1.0477778	1.376591	5.498311	1.205815
Weekday	Night	126	1	0.0001323	1.000000	NaN	NaN
Weekday	Morning	126	7306	0.9664021	1.514737	7.109786	1.090991
Weekday	Afternoon	126	6667	0.8818783	1.247388	5.039423	1.048741
Weekday	Evening	126	7535	0.9966931	1.430259	3.982906	1.204457

From the tables we can see that very few arrivals occurred during the night. For both the Fall and Winter seasons, during the weekdays there is a higher arrival rate during the mornings, and during the weekends there is a higher evening arrival rate. This could be caused by work related trips normally happening in the mornings of the weekdays, and the evening flights on the weekend as if they are on a trip in another city, they may want to spend the day in the different cities and return home at night. Another arrival rate trend we see is that the arrival rates in the morning for the winter are quite larger than the arrival rates in the morning for the fall. This could be attributed to the fact that morning flights are less likely to be delayed than other flights (https://fivethirtyeight.com/features/fly-early-arrive-on-time/ or other), and especially in a northern city like AUC (from the photo), it is likely that the winter months see snowfall related delays, so to avoid missing out on holiday vacation time in their preferred city, it is better to leave earlier for flights.

While the tables contain the average wait time for a given season, type of day, and time. We are interested in looking into the quality of service curves for our data, as well as creating quality of service curves from a regression model in order to use it for future predictions and comparing the two. In order to create the QOS curves for the regression model, we did a simple linear regression for our clusters using the formula


colour
Afternoon
Evening
Morning
Night
Weekday Fall
Weekend Fall
Weekday Winter
Weekend Winter
Quality of Service
x
0
10
20
30
0.2
0.4
0.6
0.8
1.0
0.7
0.9