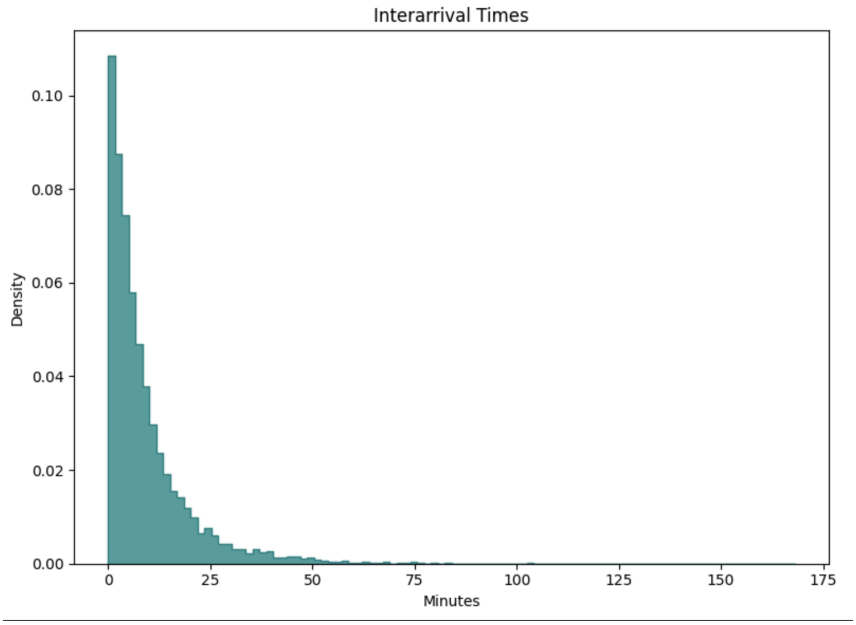
IE 305 Project
PART I - Deliverables

Group 42:
Eylül Yaltır
Sude Naz Arıcı
Tunahan Arslan
Murat Emre Aktan

In this project, timestamped order arrival data from the file Customer_Data_Clean.xlsx are analyzed to model the order arrival process for use in a discrete event simulation. The dataset was cleaned and simplified on Excel by retaining only the two columns required for arrival time analysis. Interarrival times are computed within each day to avoid distortions caused by overnight gaps.

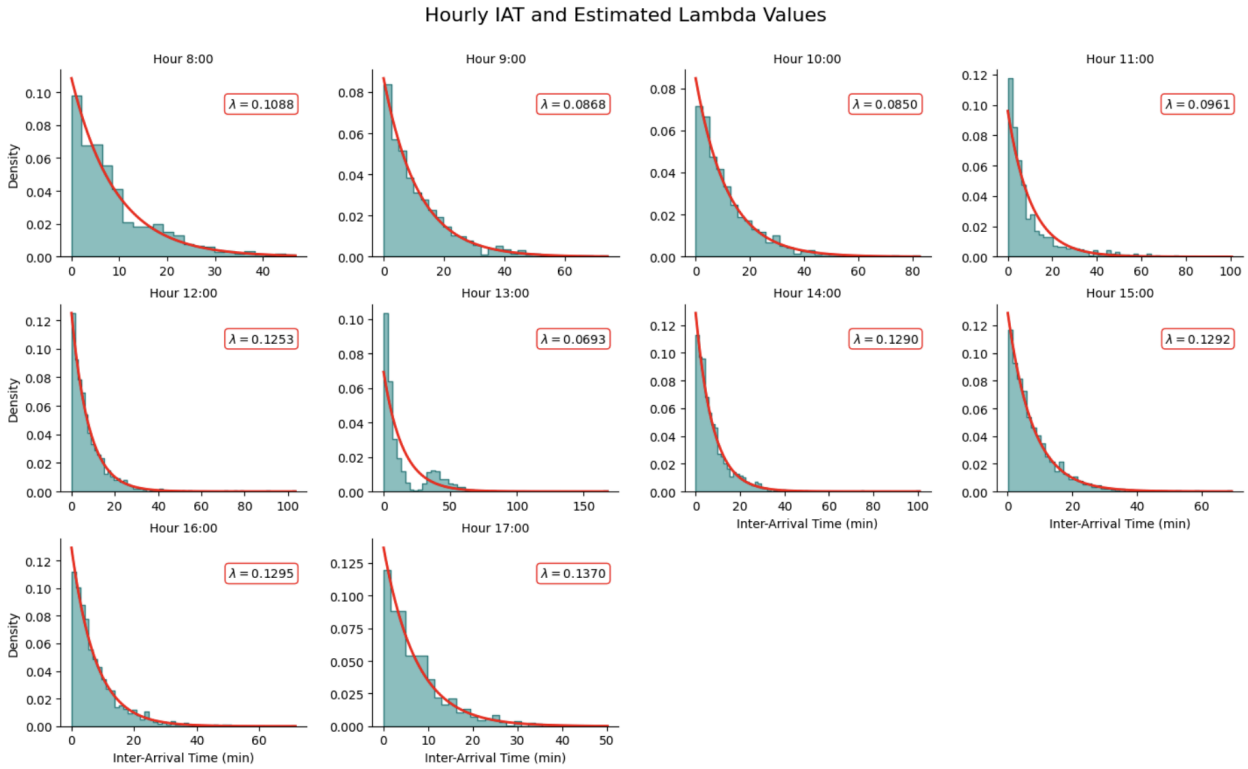
The primary objective of this analysis is to determine which mathematical distribution should be used when modelling order arrivals. In order to do this, we first created a probability density function of interarrival times in Python.

Interarrival Time:



We thought that the shape of the graph best fits an exponential function. In order to get a better understanding of the data and to create a more accurate and refined model, we decided to create several hourly clusters, and check these data independently so that we can see if the interarrival time of the demand changes throughout the day.

We tested each hour of the day in order to see if there are hourly patterns, and if each hour has its own distribution. Since the total data seemed to follow an exponential function, we looked to see if each hour has its own exponential function, and if different hours throughout the day may have similar patterns.



In this graph, we can see the distribution of interarrival times for each hour of the day. Each "cluster" of data contains all interarrival times that relate to that hour for each that that a demand has been made. As all of the hourly distributions, again, seemed to follow an exponential function, we also calculated a potential lambda value for each one of them by taking the inverse of the mean interarrival time for that hour.

Before we conducted a goodness-of-fit test for our that, we realized that throughout the day the distribution function may vary a lot, since lambda values seem to be diverse. Given this variability, modeling the entire arrival process using a single exponential distribution would be inappropriate. Instead, a time-of-day segmentation approach is adopted, where demand arrival rates are assumed to be approximately constant only within carefully selected time intervals. Based on shifts observed in the hourly arrival rates, the data are clustered into the five predefined time intervals.

Rather than assuming fixed cluster boundaries, we apply a coefficient-of-variation (CV)–based recursive splitting algorithm. This method automatically divides the hours of the day into consecutive segments where the hourly rate is statistically stable (CV below a specified threshold). Each resulting segment is interpreted as a period during which the arrival process behaves approximately like a Poisson process with exponential interarrival times.

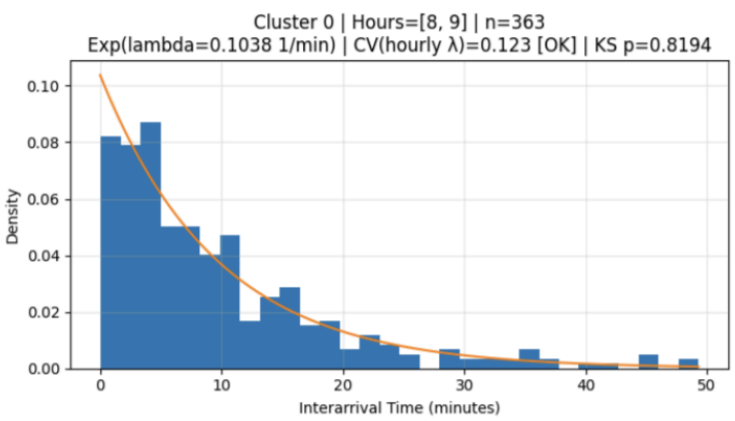
For each cluster, interarrival times are analyzed separately, and goodness-of-fit tests are applied to evaluate whether an exponential distribution provides an adequate representation. The exponential distribution yields a satisfactory fit for most clusters. However, the 13:30–14:00 interval exhibits significant deviations from exponential behavior and is better approximated by a lognormal distribution, likely due to irregular arrival patterns associated with the lunchtime period.

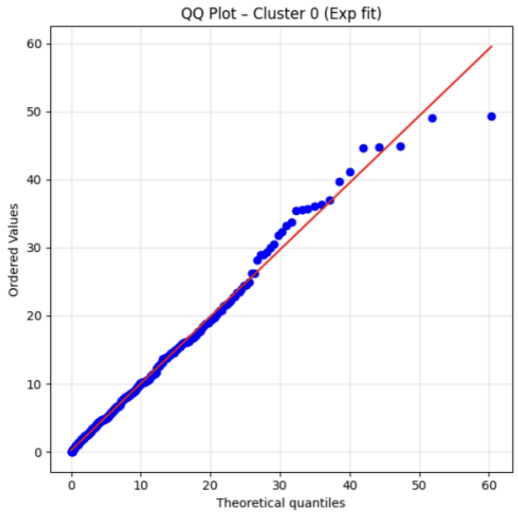
Despite this deviation, the 13:30–14:00 interval is ultimately modeled using an exponential distribution in the simulation model. This decision is made to preserve consistency with the overall arrival modeling framework and to maintain model simplicity, given the short duration and transitional nature of this time interval.

Overall, this approach captures the time-dependent nature of customer arrivals while balancing statistical accuracy and modeling practicality, ensuring that the arrival process is represented in a form suitable for simulation.

Goodness of Fit Tests

Cluster (Hr: 8-9):

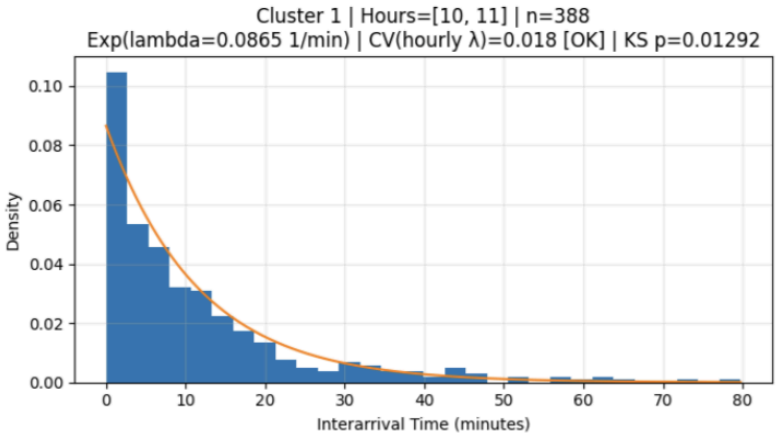


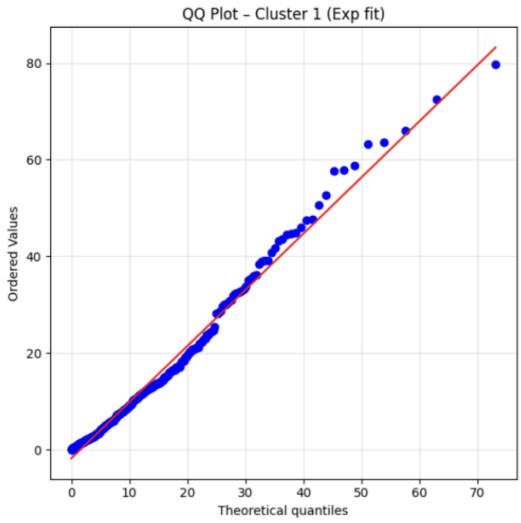


The first cluster covers the time period between 08:00 and 09:00, which represents the start of the day when arrivals begin to increase but have not yet reached their peak. When we examined the interarrival times for this cluster, the histogram showed the typical decreasing shape expected from an exponential distribution. The fitted exponential curve closely matched the observed data, and the Kolmogorov–Smirnov (KS) test returned a high p-value (0.8194), meaning that there is no evidence against the exponential assumption in this period. The QQ plot also supports this result, as most points fall closely along the reference line, indicating a good fit.

Cluster (Hr: 10-11):

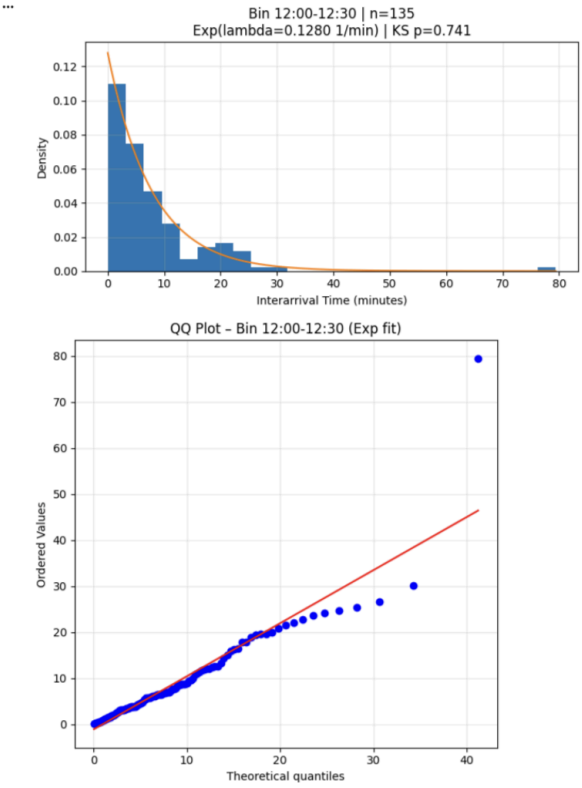
...





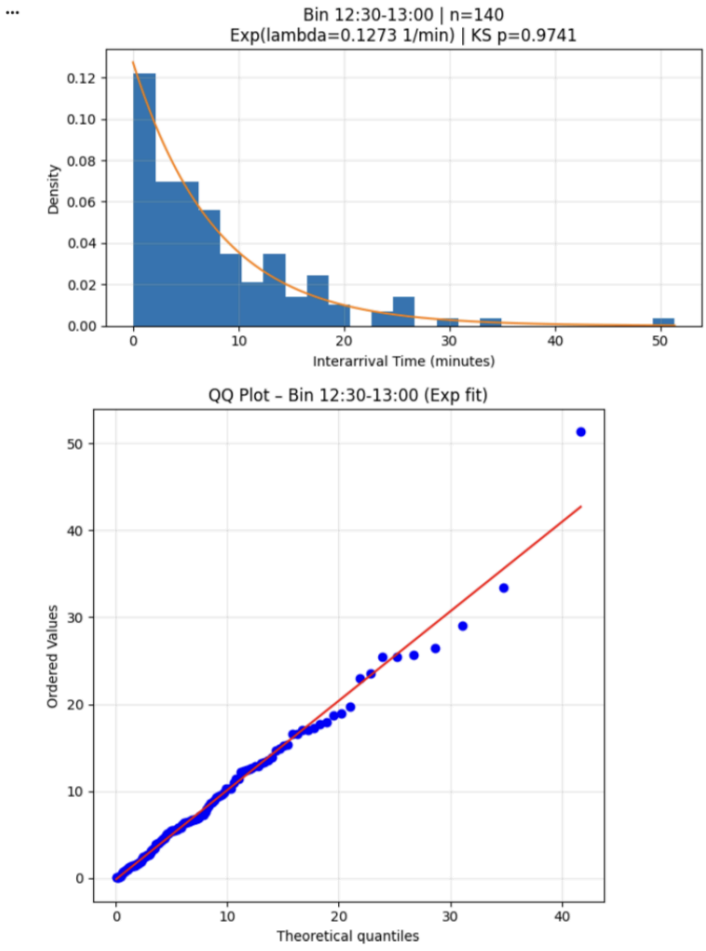
For the 10:00–11:00 period, the interarrival times show the typical decreasing shape of an exponential distribution, and the fitted curve matches the data fairly well. The KS p-value (0.0129) suggests some deviation, mainly caused by a few large interarrival times, as seen in the QQ plot's upper tail. Still, most points follow the expected exponential pattern, so modeling this cluster as exponential remains acceptable for simulation purposes.

Cluster (Hr: 12.00-12.30):



In the 12:00–12:30 period, the interarrival times align reasonably well with the exponential model. The histogram follows the expected decreasing shape, and the fitted exponential curve captures the main trend of the data. The KS p-value (0.741) indicates a good statistical fit. The QQ plot also shows that most points lie close to the reference line, with only a few larger values deviating at the upper end. Overall, this short interval behaves consistently with an exponential distribution and can be reliably modeled as such.

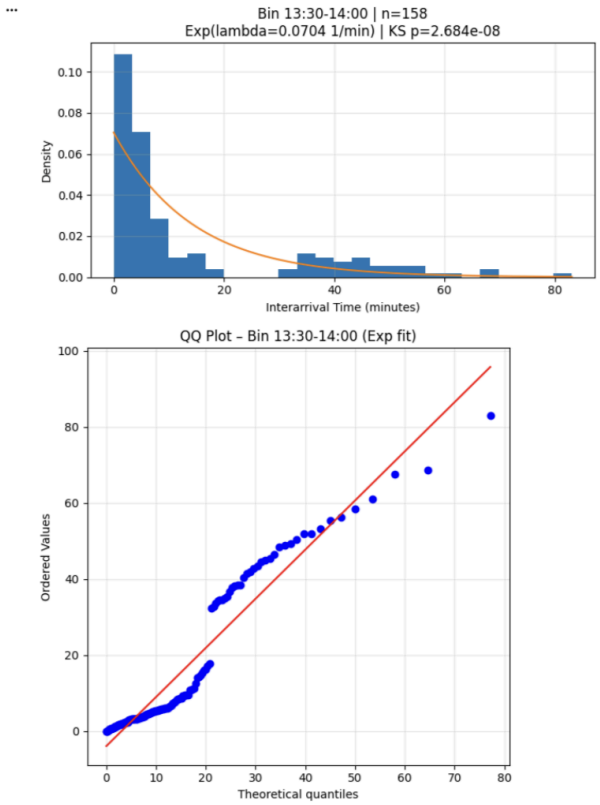
Cluster (Hr: 12.30-13.00):



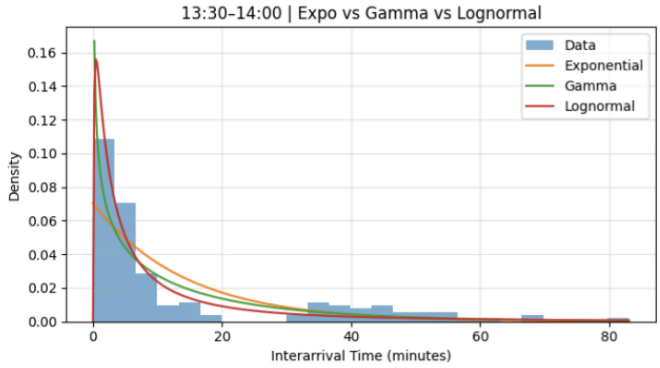
The 12:30–13:00 interval shows a strong match with the exponential distribution. The histogram follows the expected exponential decay, and the fitted curve overlays the bars closely. The KS p-value is very high (0.9741), indicating an excellent fit. The QQ plot also supports this result, with most points falling directly along the reference line except for a few larger interarrival times.

Overall, arrivals in this period behave very consistently with an exponential pattern, making this bin one of the most stable segments in the analysis.

Cluster (Hr: 13.30-14.00):

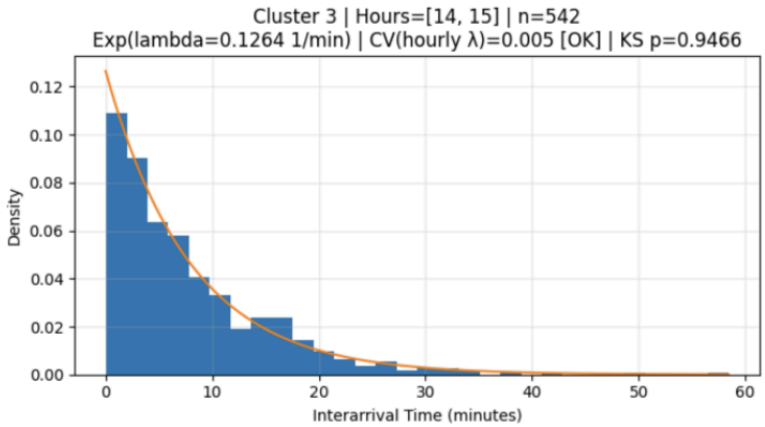


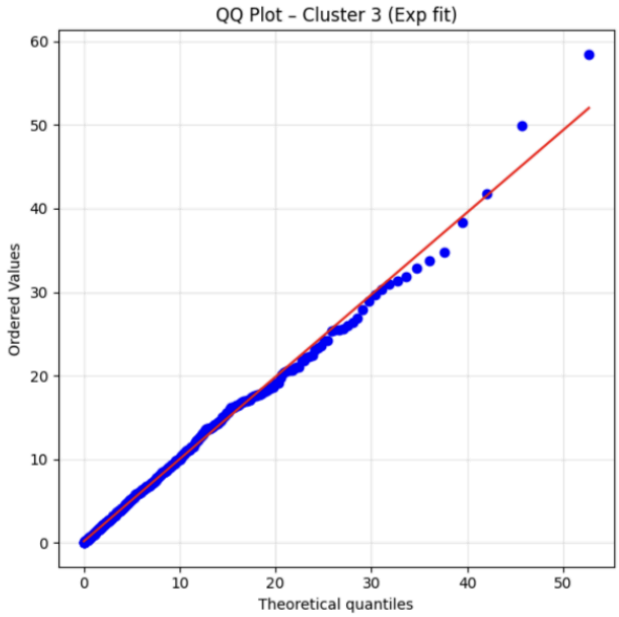
The histogram has a heavier tail than an exponential curve would predict, and the fitted exponential line does not follow the data closely. The KS p-value is extremely small, confirming a poor fit. The QQ plot also shows strong curvature, with many points falling far above the reference line.



For the 13:30–14:00 period, we compared three possible models—exponential, gamma, and lognormal—to see which one best matched the interarrival data. The histogram shows a long right tail, meaning that longer gaps between arrivals happen more often than an exponential distribution would predict. When we overlay the three fitted curves, the lognormal curve follows both the peak and the heavy tail of the data much more closely than the exponential or gamma models. This indicates that the underlying arrival behavior in this short window is more irregular and skewed, consistent with a lognormal pattern. This makes sense because this time corresponds to the lunch period, where arrivals naturally become less predictable. Although the rest of the day fits the exponential model well, lognormal provides the most accurate representation for this specific 30-minute interval.

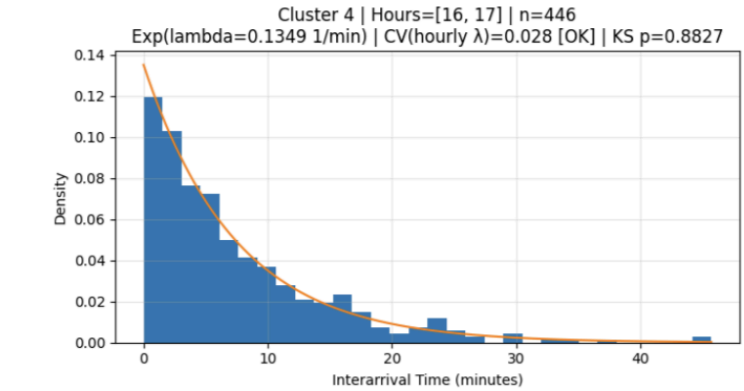
Cluster (Hr: 14-15):

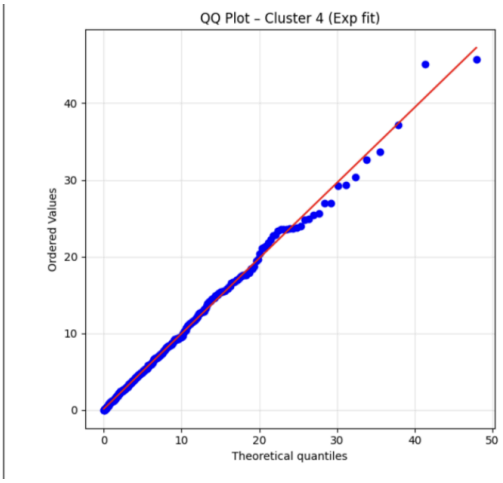




In the 14:00–15:00 time window, the interarrival times closely follow an exponential pattern. The histogram matches the expected exponential curve very well, and the KS p-value (0.9466) is high, confirming an excellent fit. The QQ plot also supports this, with nearly all points falling along the reference line except for a few large values at the tail. Overall, this cluster shows very stable arrival behavior, making the exponential distribution a strong and reliable model for this period.

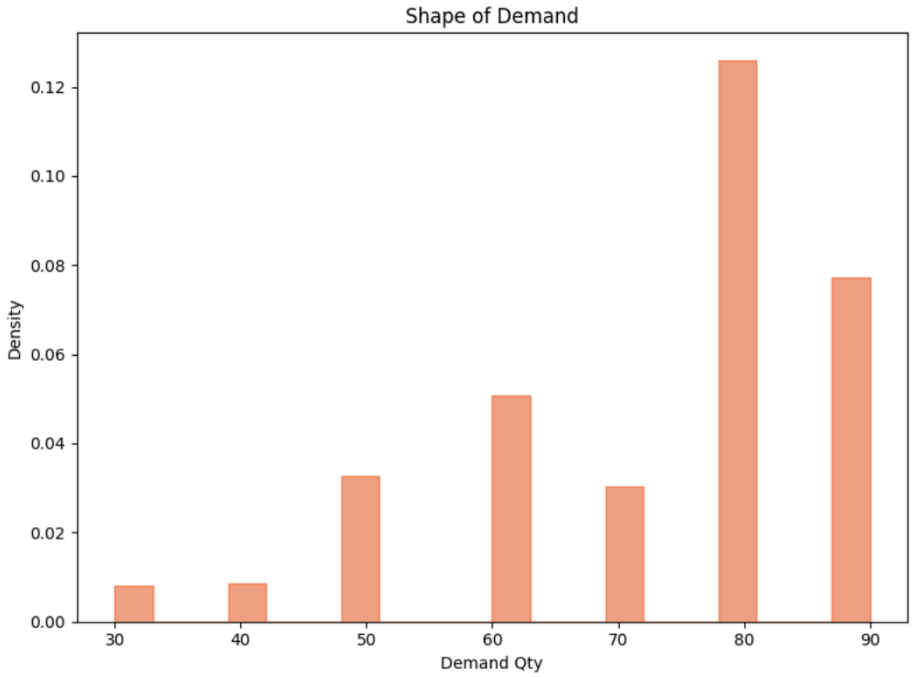
Cluster (Hr:16-17):





For the 16:00–17:00 period, the interarrival data fits the exponential distribution very well. The histogram and the fitted exponential curve align closely, and the KS p-value (0.8827) shows no evidence against the exponential model. The QQ plot also supports this, with most points lying almost exactly along the reference line, except for a few larger values at the tail. Overall, this cluster displays stable and consistent arrival behavior, making the exponential distribution an excellent model for this time window.

Demand



For the demand data we determined that the variable is discrete, meaning it takes on specific, fixed integer values rather than a continuous range. Because of this, fitting a theoretical continuous curve can be misleading, as it might predict impossible decimal values (e.g., 42.5). Therefore we thought that the most accurate approach is to use the Empirical Distribution, which simply calculates the exact probability of each demand value occurring based on its historical frequency. Here is the table of frequencies within the data.

Demand Value	Probability (PDF)	Cumulative Probability (CDF)
30	2.55%	0.0255
40	2.72%	0.0526
50	10.36%	0.1562

60	15.01%	0.3063
70	9.46%	0.4010
80	37.12%	0.7722
90	22.78%	1.0000