به نام خدا

تمرین شماره‌ی۳

زبان‌های شبیه سازی

احمد امامی

۹۹۲۰۷۵۲۱

** تمام بررسی‌ها و نتایج ارائه شده در این تمرین به کمک پایتون انجام شده است و کد آن به پیوست ضمیمه شده است.

فرضیات

در این تمرین داده‌های تونژ کشتی GRT، طول کشتی Length، تعداد کانتینرها TEU و زمان بین ورود Interarrival time را به عنوان ورودی سیستم شبیه‌سازی در نظر گرفتیم. در ارتباط با ویژگی setup time که در کلاس نیز مورد بحث واقع شد، تصمیم گرفتیم که به عنوان خروجی سیستم در نظر گرفته شود. زیرا عوامل مختلفی در سیستم می‌تواند در زمان آماده‌سازی کشتی دخیل باشد و این موضوع سبب می‌شود که نیاز به تحلیل نتایج خروجی داشته باشیم. در نتیجه در این تمرین ۴ ویژگی عنوان شده در بالا مورد بررسی قرار می‌گیرند.

هم‌چنین در این تمرین مقادیر مشخصات فوق را به صورت جداگانه برای هر کشتی مورد بررسی قرار داده‌ایم.

پیش‌پردازش داده‌ها Data Cleaning

پیش از پاسخ به پرسش های تمرین بهتر است دیتاست مربوطه را بررسی کرده و در صورت وجود داده‌ی null و یا داده‌های پرت آن‌ها را حذف کنیم. زیرا در فرایند فیت کردن توزیع به هرکدام از مشخصه‌ها، در صورت وجود داده‌های پرت با مشکل مواجه می‌شویم.

ابتدا ستون No از دیتاست را حذف می‌کنیم زیرا نیازی به آن نخواهیم داشت. ۵ سطر ابتدایی دیتاست جدید را مشاهده می‌کنید.

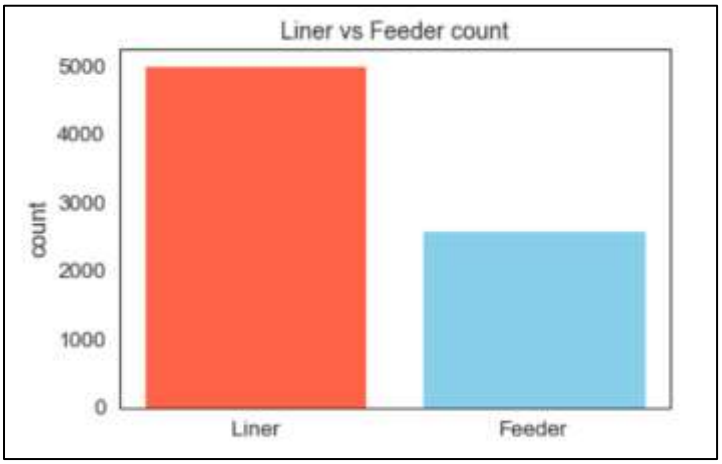
	GRT	Length	TerminalName	TEU	VoyageType	Interval	EarlinessTardiness	ST	OT	UT
1	17618.0	201	1	1861	Feeder	18.23	13	4.00	31.40	3.27
2	16694.0	174	1	1223	Liner	8.90	64	2.72	46.50	5.78
3	29873.0	210	1	605	Liner	30.38	0	1.65	25.85	4.08
4	15670.0	168	1	1502	Liner	12.00	7	4.38	56.13	5.82
5	16100.0	170	1	268	Liner	8.17	15	0.77	7.63	1.60

داده‌های null را بررسی می‌کنیم:

```
1  portdata.isnull().sum()
✓ 0.3s
GRT                0
Length             0
TerminalName       0
TEU                0
VoyageType         0
Interval           0
EarlinessTardiness 0
ST                 0
OT                 0
UT                 0
dtype: int64
```

همان‌طور که می‌بینیم در هیچ یک از ستون‌های دیتاست داده‌ی null مشاهده نمی‌شود و از این نظر مشکلی در دیتاست مشاهده نمی‌شود.

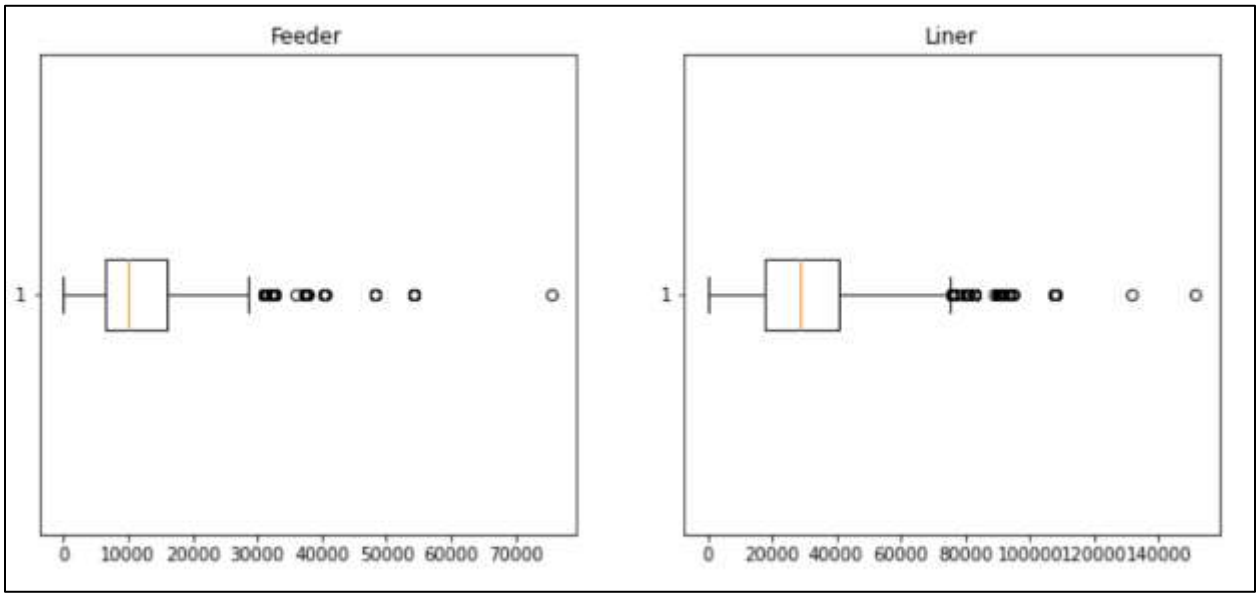
در مرحله‌ی بعدی بهتر است که داده‌هایمان را به دو دسته تقسیم کنیم. یک قسمت مربوط به کشتی‌های لاینر و دیگری مربوط به کشتی‌های فیدر. از آنجایی که مشخصات مرتبط با هر کشتی مختص خودش است، بهتر است آن‌ها را به صورت جداگانه و در دو دیتاست مختلف مورد بررسی قرار دهیم. همان طور که در تصویر زیر می‌بینیم حدودا ۵ هزار کشتی لاینر و ۲۵۰۰ عدد نیز فیدر هستند که به صورت مجزا مورد بررسی قرار می‌دهیم.



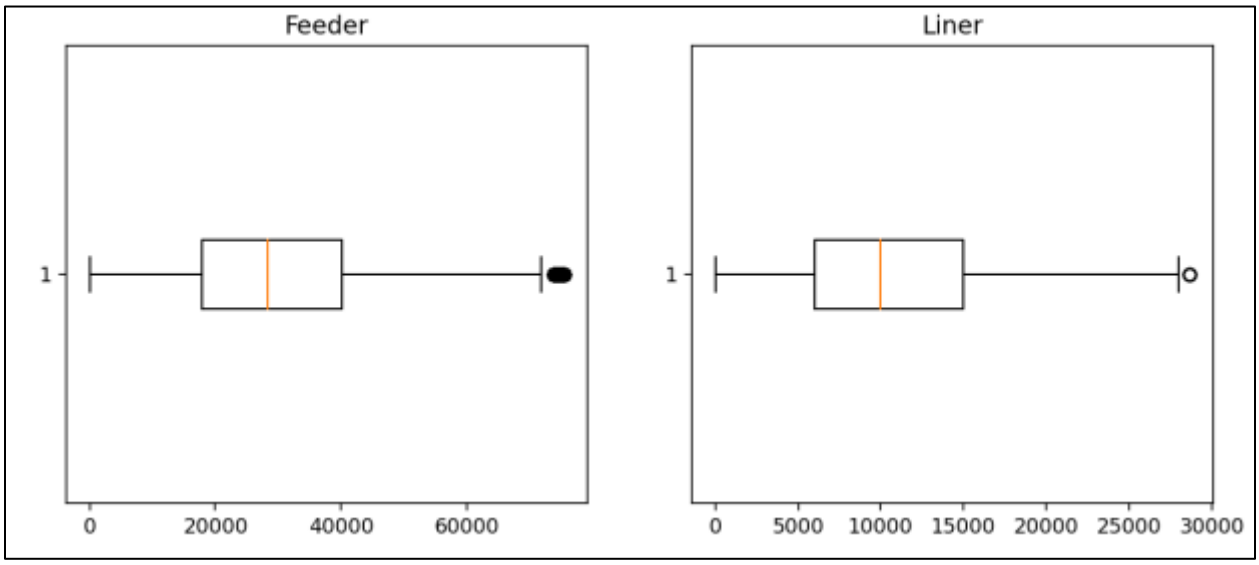
حذف داده‌های پرت

برای حذف داده‌های پرت از نمودار جعبه‌ای (boxplot) کمک می‌گیریم. برای هر مشخصه این نمودار را رسم می‌کنیم و داده‌های بسیار بزرگ و غیر معمول را حذف می‌نماییم.

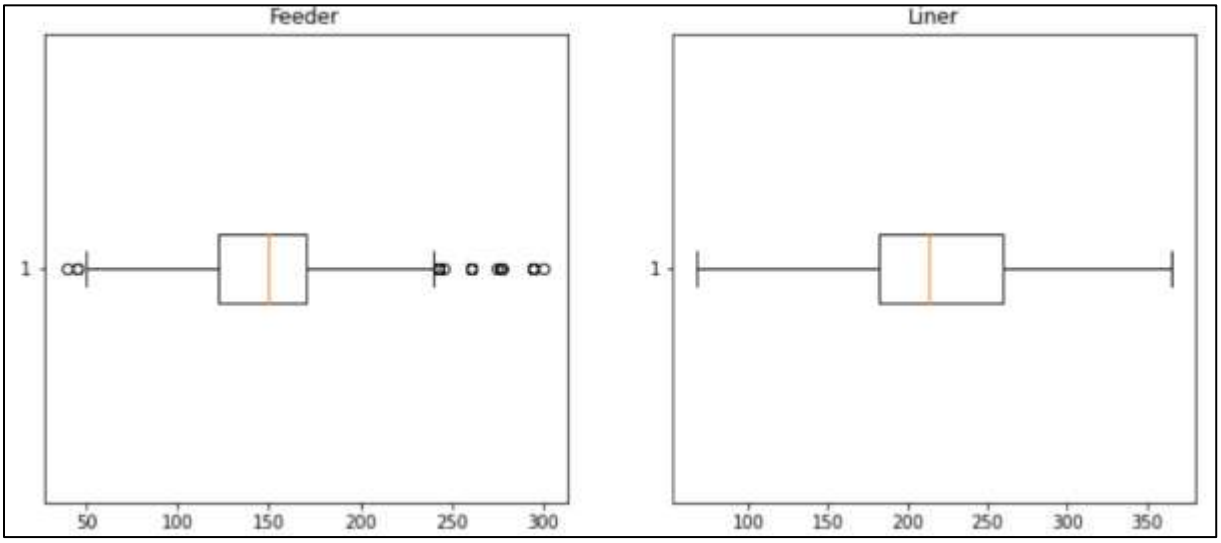
GRT



پس از حذف داده‌های پرت نمودار باکس‌پلات شاخصGRT به شکل زیر خواهد بود:

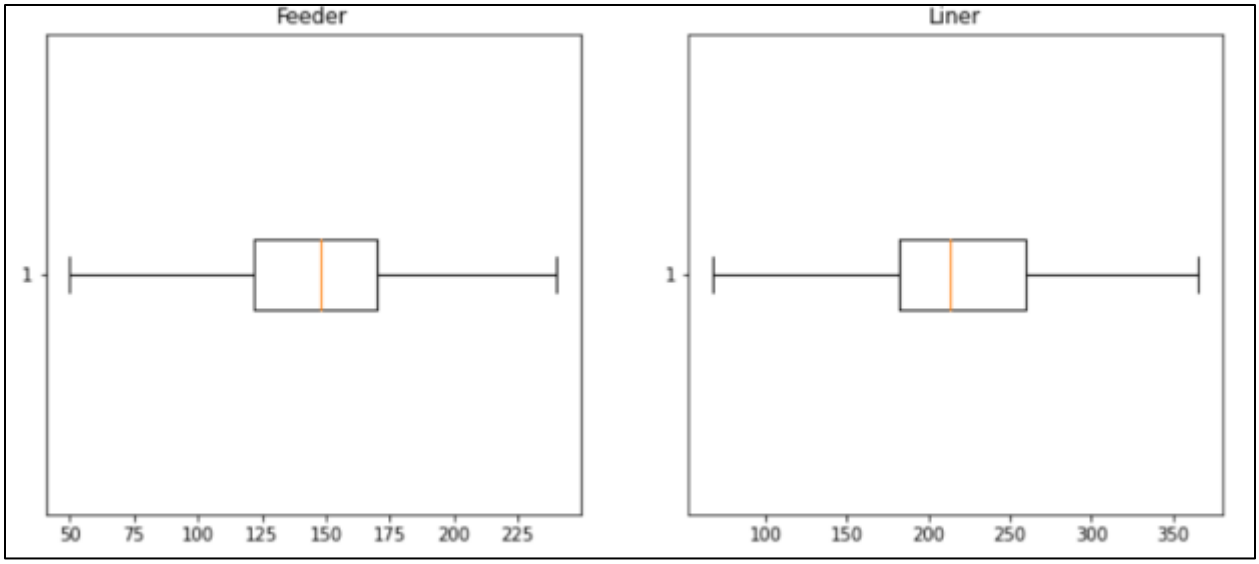


Length

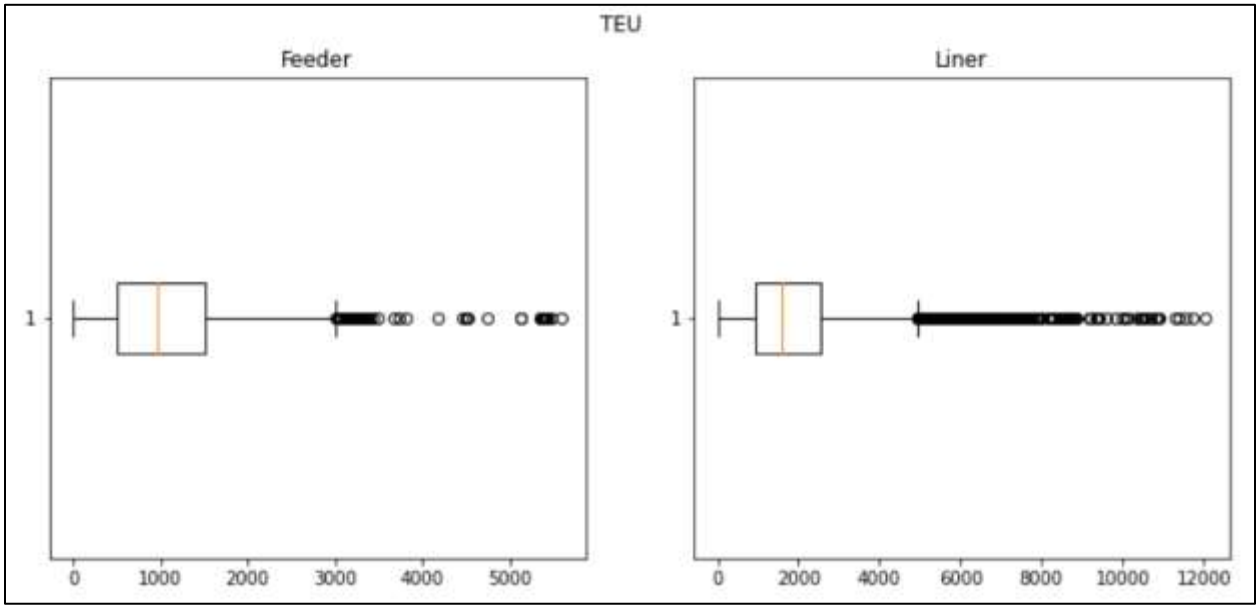


برای طول کشتی‌ها وضعیت بهتری را شاهد هستیم و برای کشتی‌های Liner داده‌ی پرتی مشاهده نمی‌شود. برای کشتی‌های Feeder تعدادی مشاهده outlier داریم که تصمیم گرفتیم که آن‌ها را حذف بنماییم.

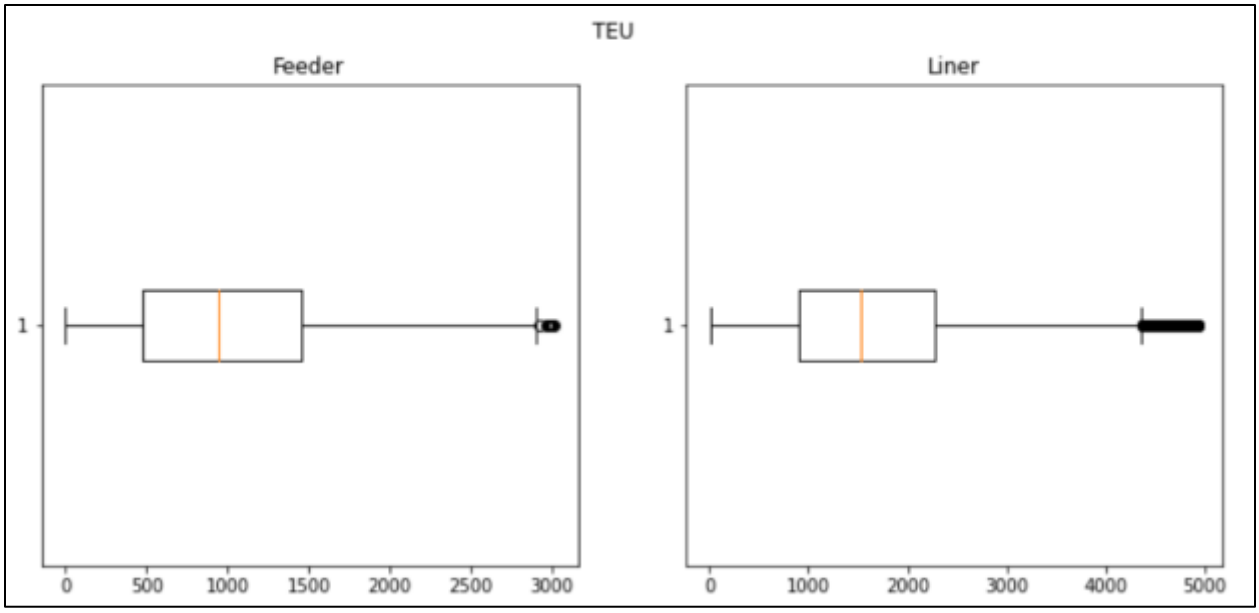
باکس‌پلات داده‌های جدید طول کشتی به شکل زیر است:



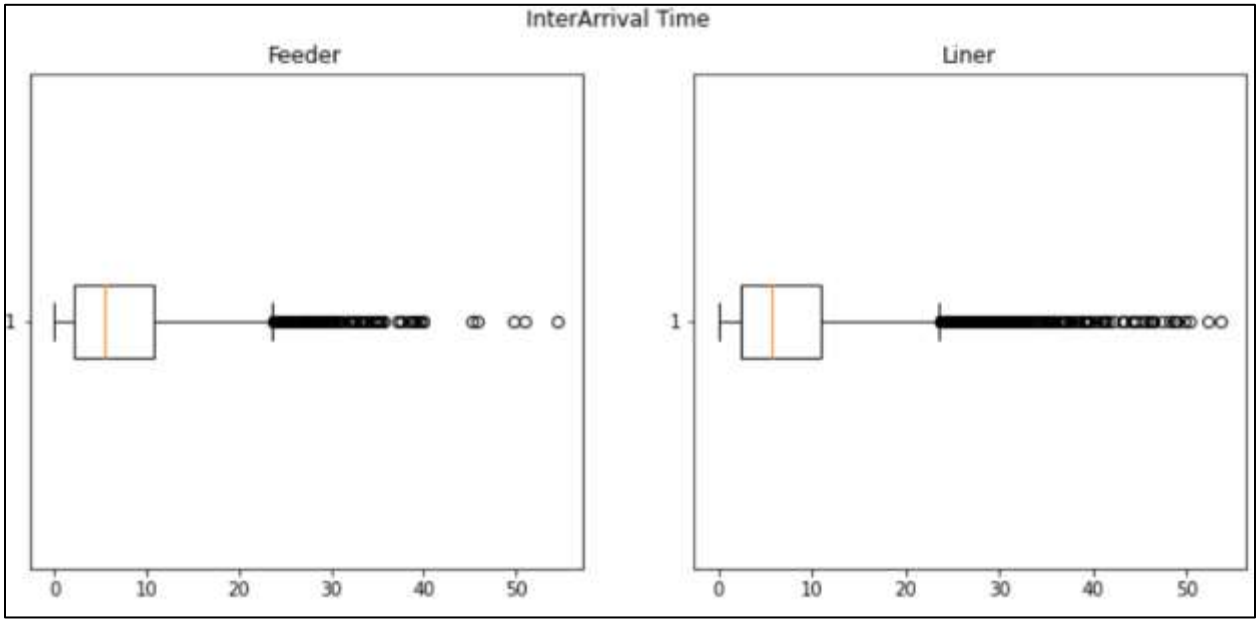
TEU

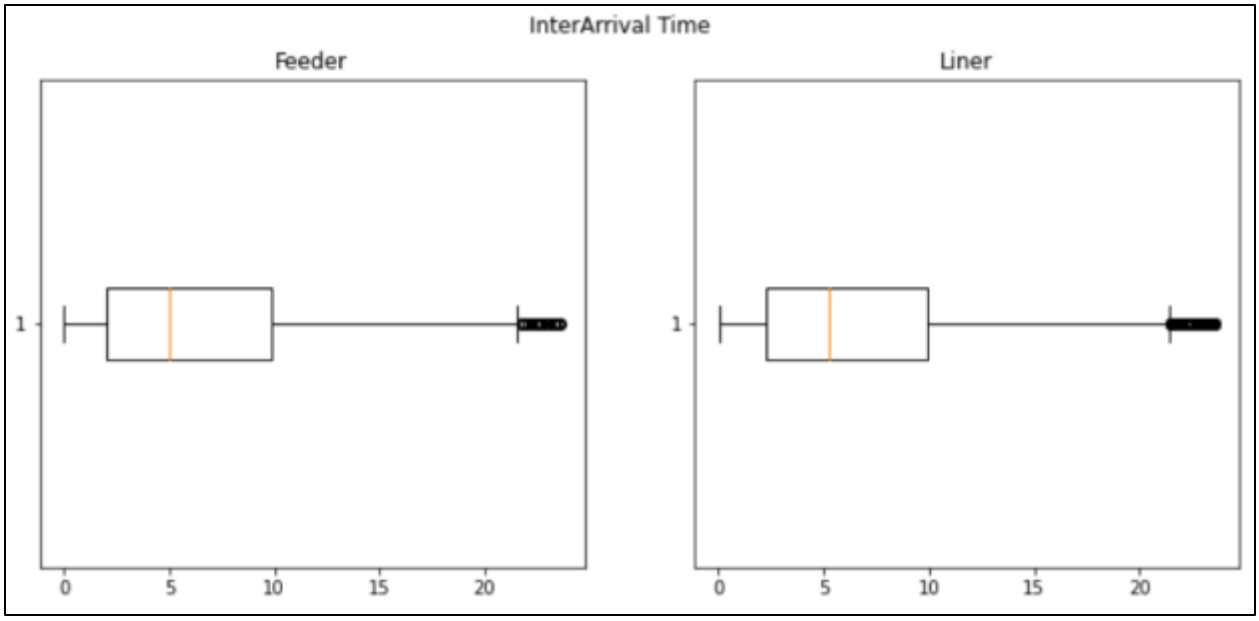


همان‌طور که مشاهده می‌شود واضح است که داده‌ها توزیع چندان مناسبی ندارند و بهتر است داده‌های پرت را حذف کنیم. نمودار باکس‌پلات بروز شده در شکل زیر قابل مشاهده است.



Interarrival time





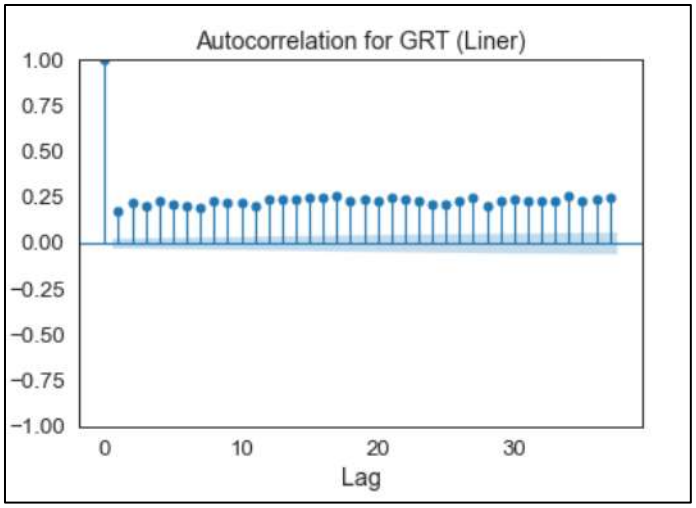
سوال ۱

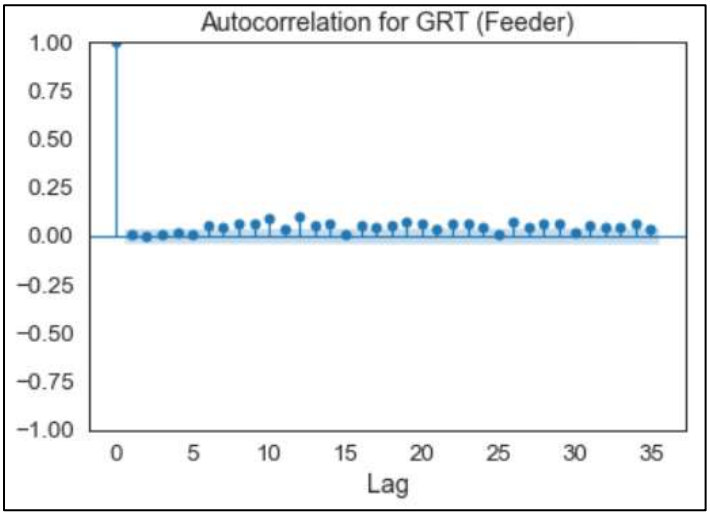
بررسی فرض iid بودن

برای بررسی iid بودن باید چک شود که آیا داده‌های مربوطه در هر ستون به یکدیگر مرتبط هستند یا خیر. به بیان دیگر آیا خوهمبستگی autocorrelation میان داده‌ها دیده می‌شود یا نه. اگر فرض iid بودن برقرار باشد آنگاه داده‌های ما مانند یک نویز رفتار می‌کنند و بین هیچ لگی همبستگی‌ای وجود نخواهد داشت. یکی از بهترین راه‌ها برای بررسی این موضوع استفاده از نمودار خودهمبستگی (ACF) می‌باشد. در صورت مشاهده‌ی یک لگ قابل توجه در این نمودار میتوان نتیجه گرفت که داده‌ها دارای خودهمبستگی هستند. نتایج زیر را برای مشخصه‌های مختلف دیتاست مربوطه مشاهده می‌کنید:

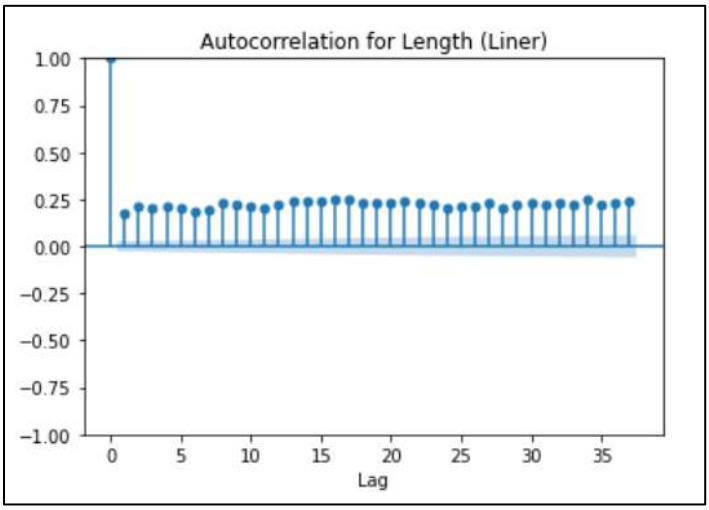
تونژ ناخالص کشتی GRT

همان‌طور که در تصاویر زیر مشاهده می‌شود مقدار همبستگی میان تونژ ناخالص در کشتی‌های لاینر و فیدر دارای خودهمبستگی است و در لگ‌های متعددی مقادیر قابل توجه مشاهده می‌شود. در نتیجه فرض i.i.d بودن در ارتباط با مشخصه‌ی تونژ ناخالص کشتی قابل قبول نمی‌باشد.

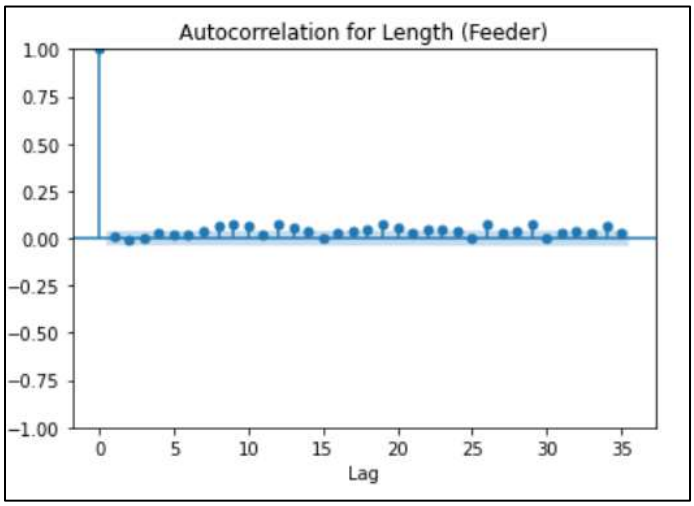




طول کشتی‌ها Length

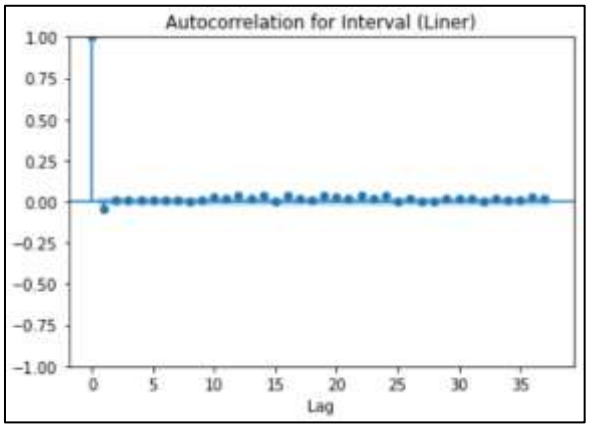


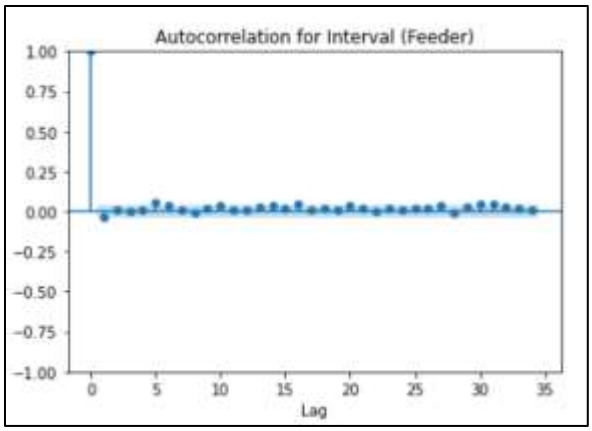
برای کشتی‌های لاینر واضح است که طول کشتی دارای خودهمبستگی است و فرض i.i.d بودن برقرار نیست.



برای کشتی‌های Feeder مشاهده می‌کنیم که طول کشتی‌ها نزدیک به iid است و میتوان فرض کرد استقلال خطی میان آن‌ها برقرار است.

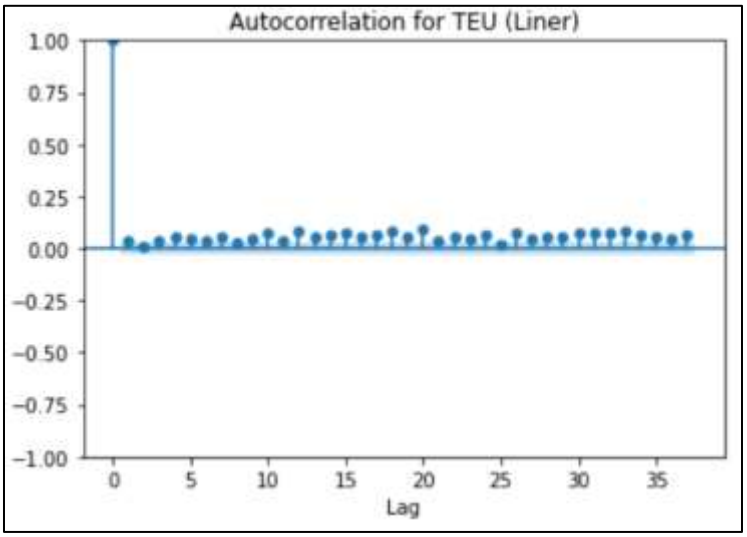
زمان بین ورود (Interarrival time)

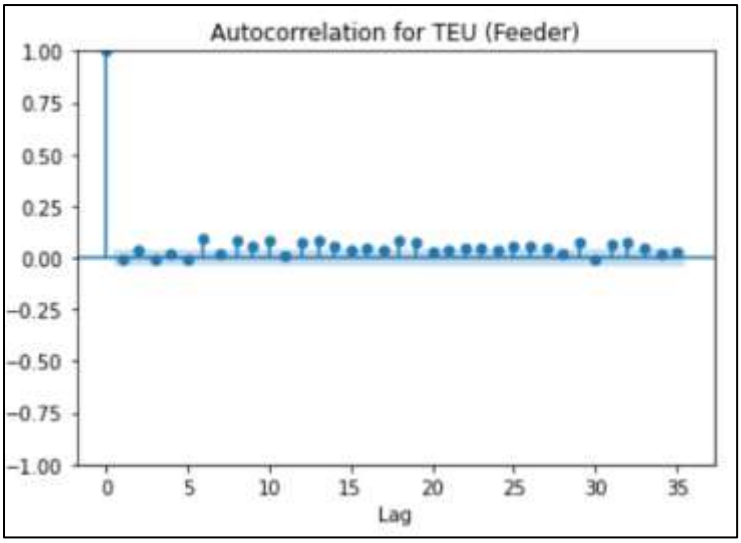




برای زمان بین ورود هر دو کشتی مشاهده می‌کنیم که شرط iid بودن برقرار است.

تعداد کانتینرهای کشتی TEU



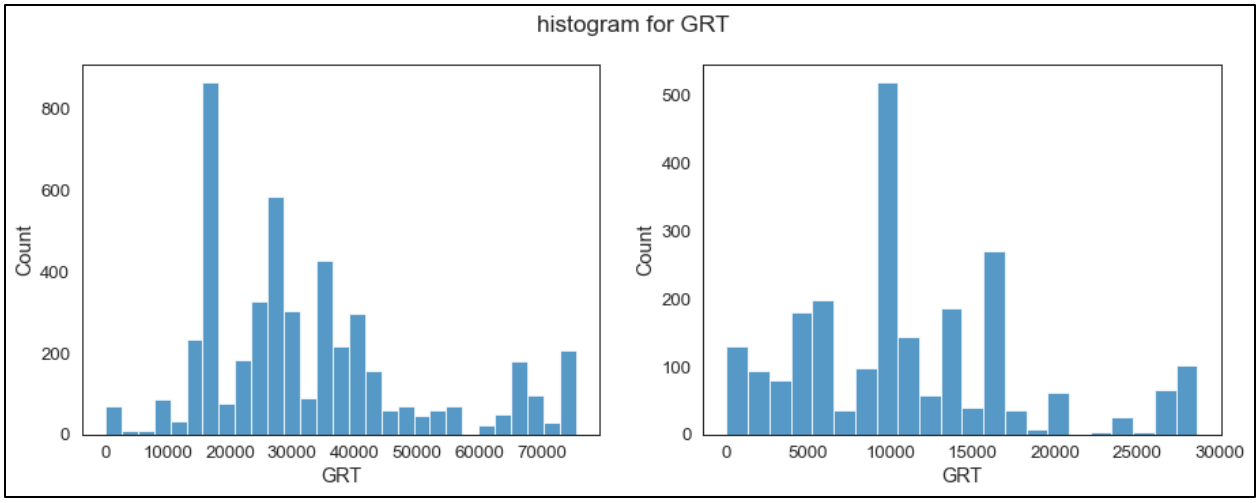


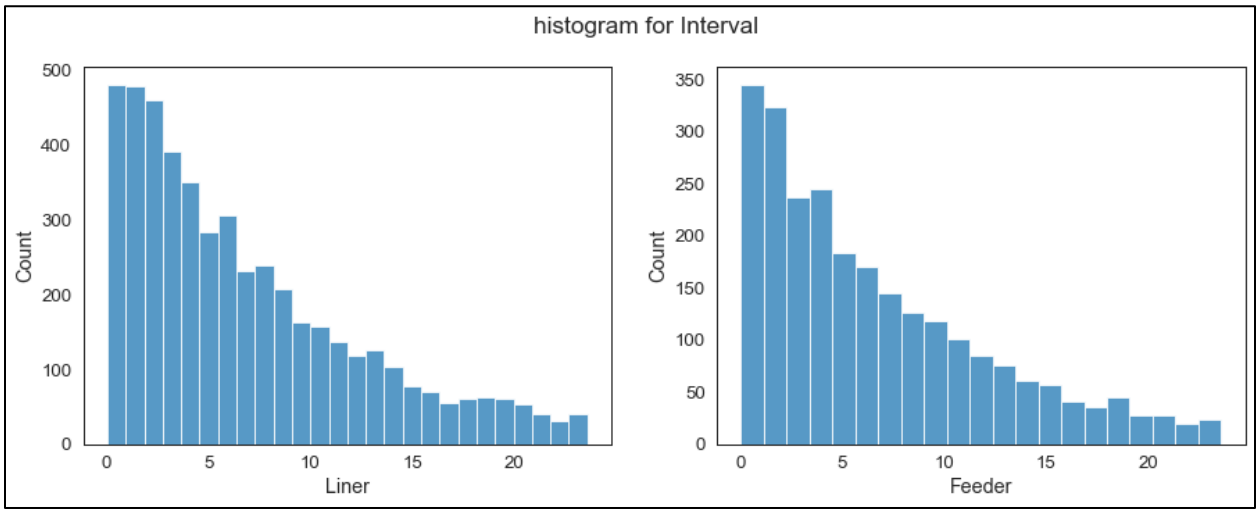
بار کشتی‌های لاینر و فیدر نیز تا حد خوبی به iid بودن نزدیک است و از این رو میتوان فرض کرد از توزیع یکسانی برخوردار هستند.

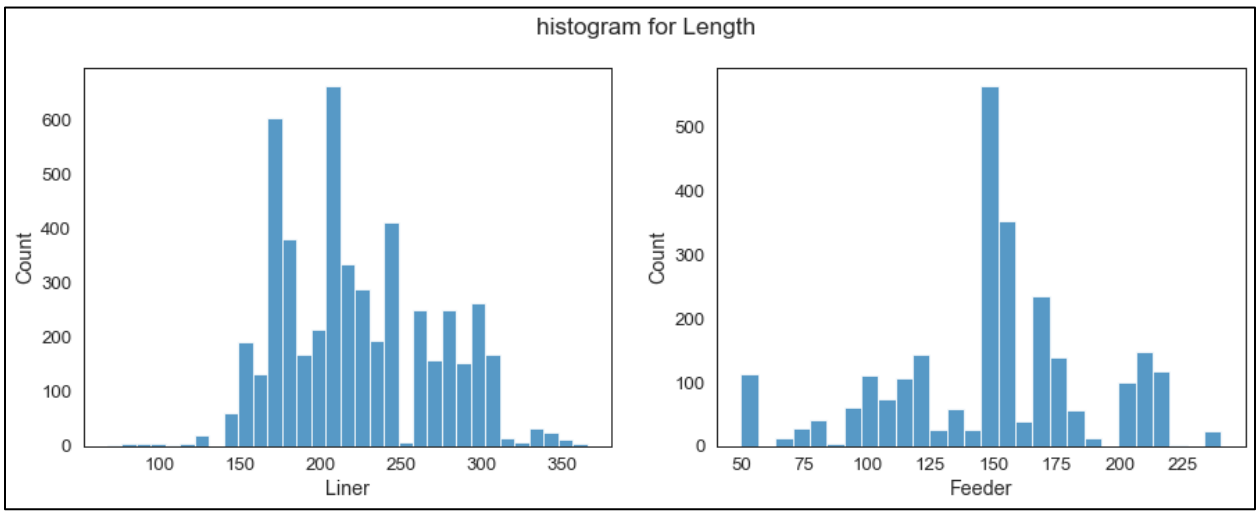
در ادامه نمودار هیستوگرام مشخصه‌های فوق را رسم می‌کنیم و توزیع‌های احتمالی را برای هر یک بیان می‌کنیم.

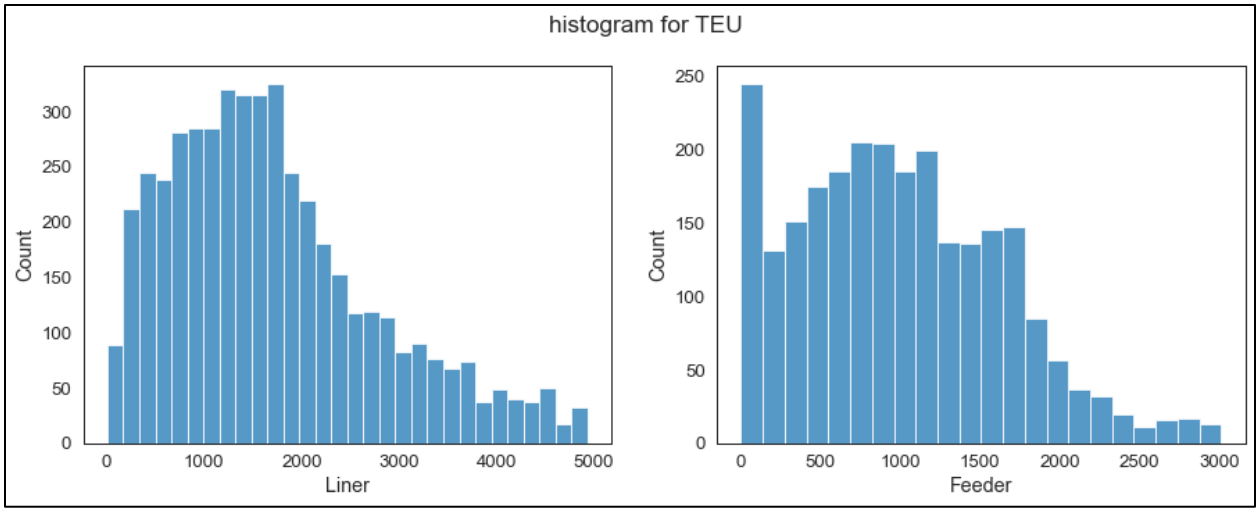
بررسی هیستوگرام داده‌های ورودی

برای رسم هیستوگرام‌ها نیز مجددا از پایتون استفاده کردیم و نتایج را در شکل‌های زیر مشاهده می‌نمایید.


histogram for GRT
Count
GRT
0
200
400
600
800
0
10000
20000
30000
40000
50000
60000
70000
Count
GRT
0
100
200
300
400
500
0
5000
10000
15000
20000
25000
30000

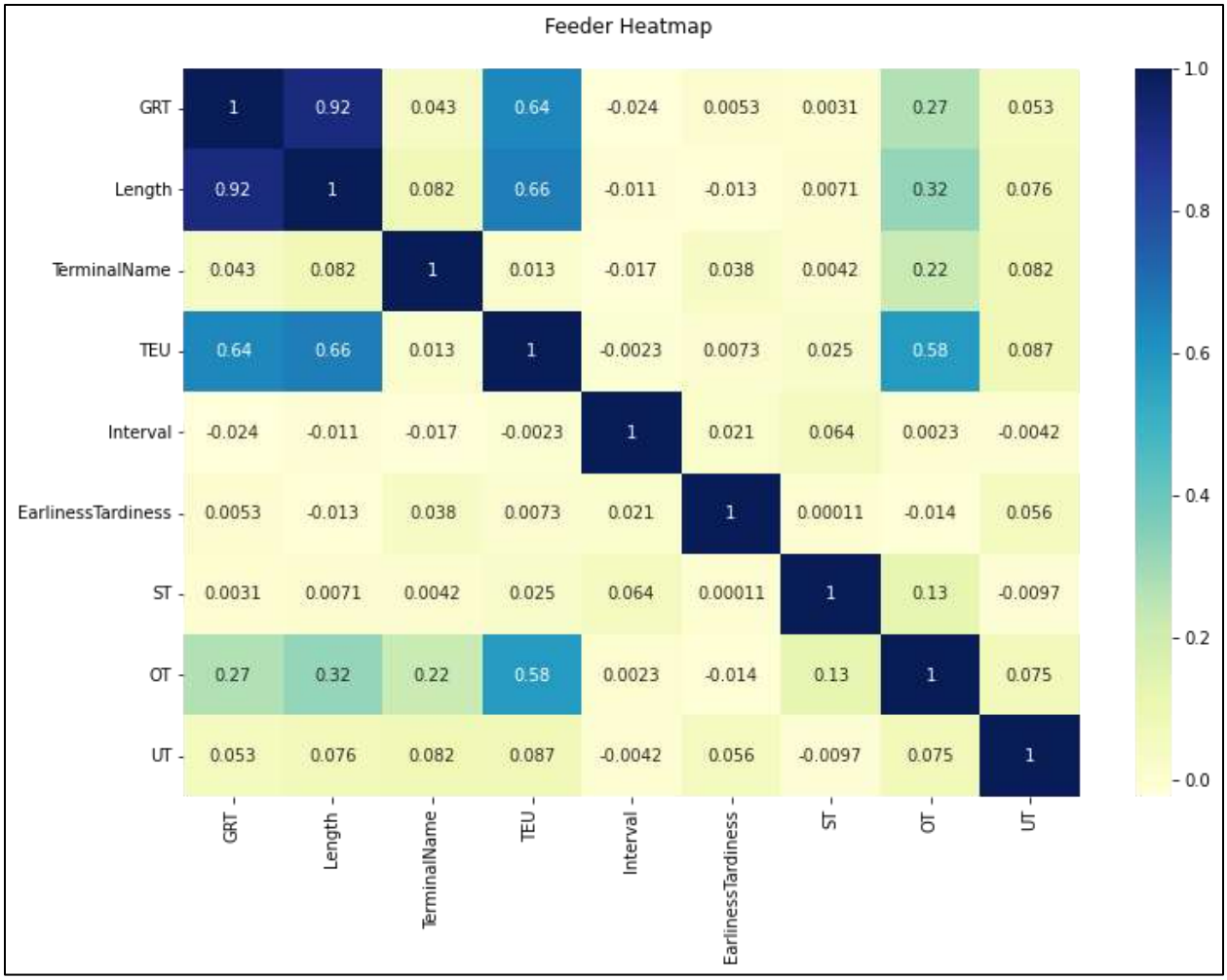

histogram for Interval
Count
Liner
0
100
200
300
400
500
0
5
10
15
20
Count
Feeder
0
50
100
150
200
250
300
350
0
5
10
15
20


histogram for Length
Count
Liner
0
100
200
300
400
500
600
100
150
200
250
300
350
Count
Feeder
0
100
200
300
400
500
50
75
100
125
150
175
200
225



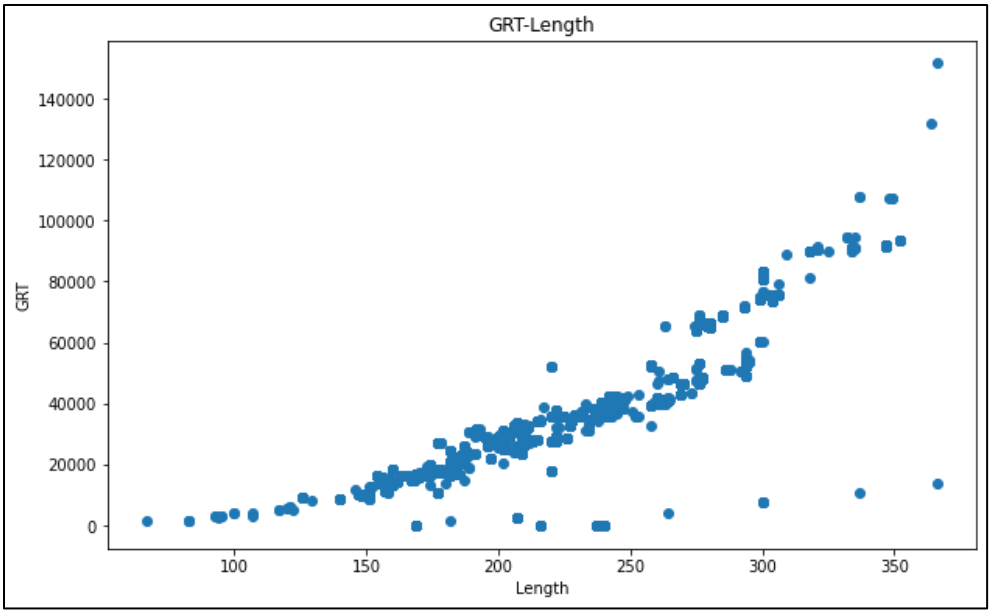
همبستگی میان ویژگی‌ها

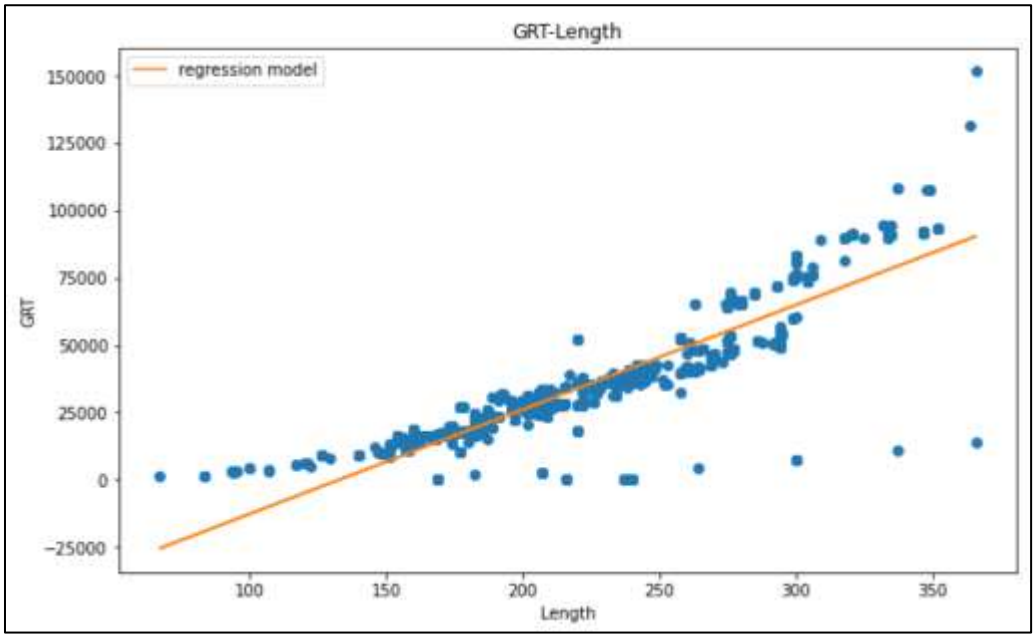
برای بهتر نشان دادن ارتباط و همبستگی میان مشخصه‌های مختلف دیتاست از ماتریس همبستگی استفاده کردیم. نتایج حاصله برای کشتی‌های Liner و Feeder به شکل زیر است.



همان طور که مشاهده می‌شود دو ویژگی TEU و GRT همبستگی بسیار زیادی با یکدیگر دارند. میتوان برای درک بیشتر رگرسیون خطی بین این دو مقدار را نیز به دست آورد. پیش از آن نمودار تونژ کشتی بر حسب طول آن را در رسم می‌کنیم. همان طور که در نمودار مشاهده می‌شود و انتظار داشتیم یک همبستگی خطی میان این دو ویژگی دیده می‌شود.

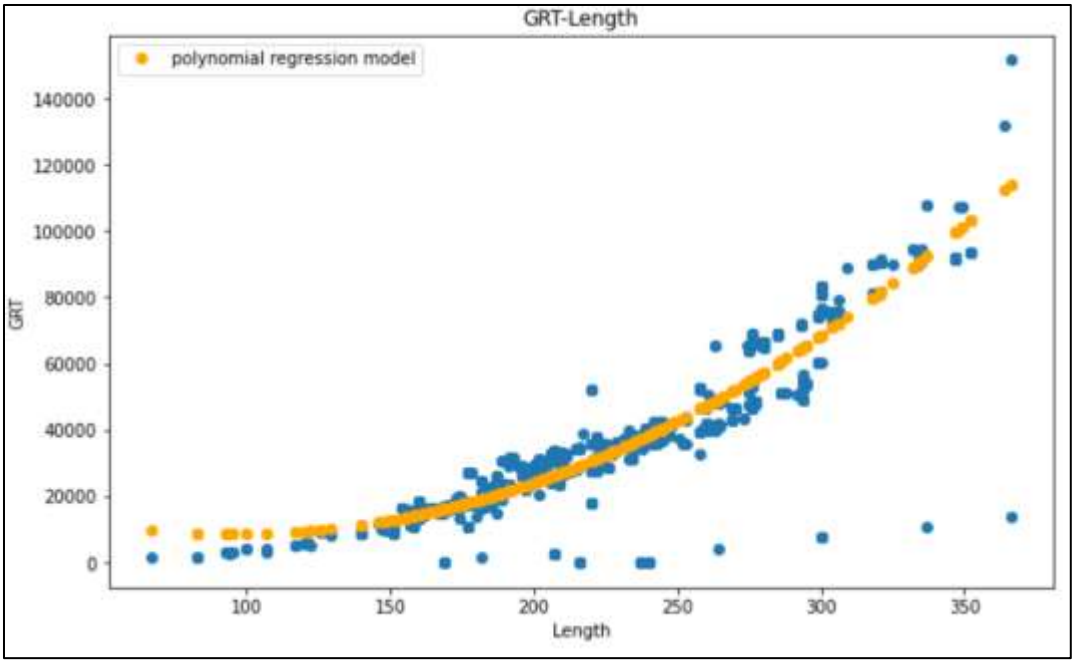
زمانی که دو یا چند ویژگی در دیتاست مورد بررسی‌مان دارای همبستگی خطی هستند بهتر است از رگرسیون کمک بگیریم و داده‌ی دیگر را بر حسب دیگری بیان کنیم. در این صورت میتوان یکی از آن‌ها را از مدل شبیه‌سازی حذف نمود. در ادامه مدل رگرسیونی میان این دو ویژگی را بیان می‌کنیم.





با توجه به خروجی نرم افزار معادله‌ی خط به صورت $387x - 51516$ می‌باشد

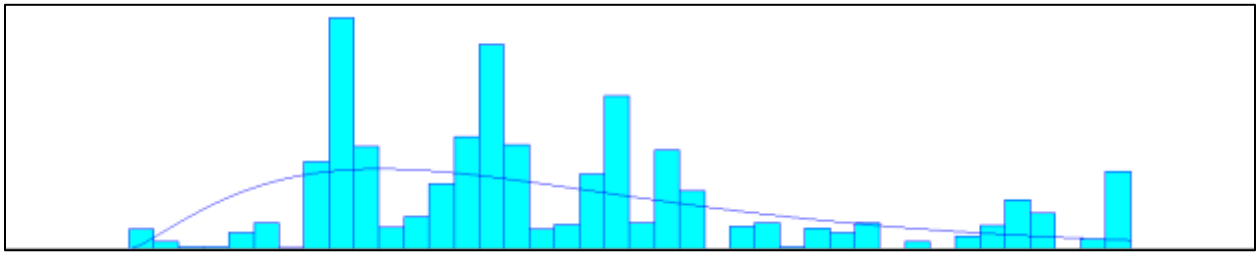
هم‌چنین میتوان رگرسیون‌های مرتبه‌ی بالاتر را نیز پیاده‌سازی نمود. به عنوان مثال در شکل زیر یک رگرسیون درجه دوم را فیت نموده‌ایم که نتیجه را مشاهده می‌کنید:



فیت کردن داده‌ها و آزمون‌های برازش نیکویی

برای این قسمت از تمرین ابتدا داده‌های پیش‌پردازش شده را در پایتون فراخوانی می‌کنیم و به تفکیک نوع کشتی در فایل txt مجزا قرار می‌دهیم. سپس به کمک ابزار input analyzer در arena بهترین توزیع را به آن فیت می‌کنیم و نتایج آزمون‌های برازش نیکویی را مشاهده می‌کنیم. نتایج حاصله به تفکیک در ادامه آورده شده اند:

GRT-Liner

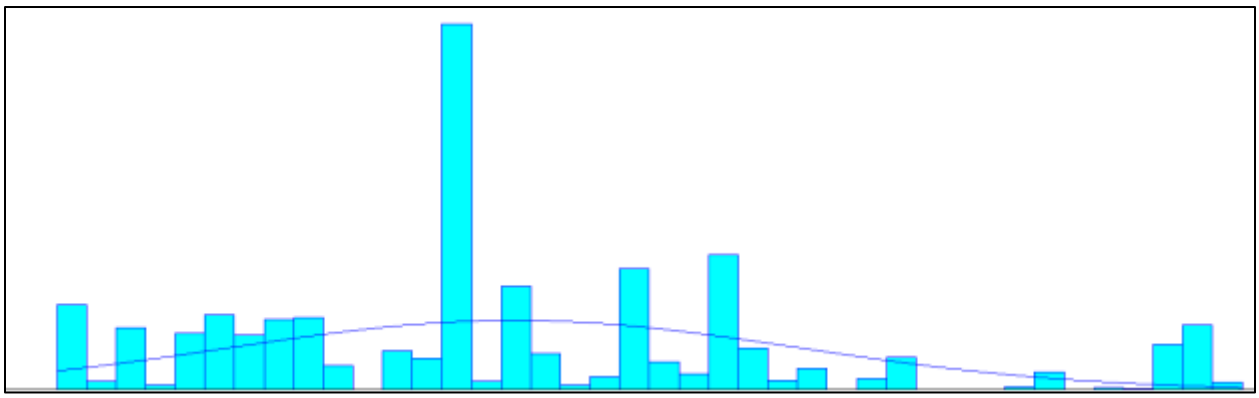


```
        Distribution Summary

Distribution:   Gamma
Expression:     18 + GAMM(1.38e+04, 2.37)
Square Error:   0.026513

Chi Square Test
  Number of intervals   = 40
  Degrees of freedom    = 37
  Test Statistic        = 5.25e+03
  Corresponding p-value < 0.005

Kolmogorov-Smirnov Test
  Test Statistic        = 0.129
  Corresponding p-value < 0.01
```

```
        Data Summary

Number of Data Points   = 4822
Min Data Value          = 18
Max Data Value          = 7.55e+04
Sample Mean             = 3.27e+04
Sample Std Dev          = 1.74e+04

        Histogram Summary

Histogram Range         = 18 to 7.55e+04
Number of Intervals     = 40
```

همان‌طور که در نتایج مشاهده می‌شود فرض ۰ در هر دو تست برازش رد می‌شوند و بدین معنی است که توزیع مورد نظر فیت مناسبی برای داده‌های ما نیست. دلیل اصلی این اتفاق iid نبودن داده‌ها است.

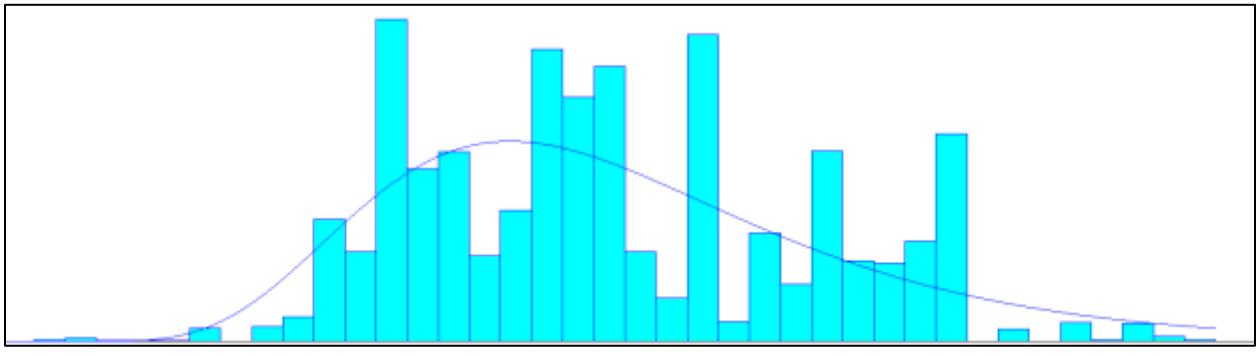
GRT-Feeder



```
        Distribution Summary

Distribution:    Normal
Expression:      NORM(1.11e+04, 6.88e+03)
Square Error:    0.049354

Chi Square Test
  Number of intervals   = 37
  Degrees of freedom    = 34
  Test Statistic        = 4.15e+03
  Corresponding p-value < 0.005

Kolmogorov-Smirnov Test
  Test Statistic        = 0.184
  Corresponding p-value < 0.01
```

```
          Data Summary

Number of Data Points    = 2332
Min Data Value           = 31
Max Data Value           = 2.87e+04
Sample Mean              = 1.11e+04
Sample Std Dev           = 6.88e+03

        Histogram Summary

Histogram Range          = 31 to 2.87e+04
Number of Intervals      = 40
```

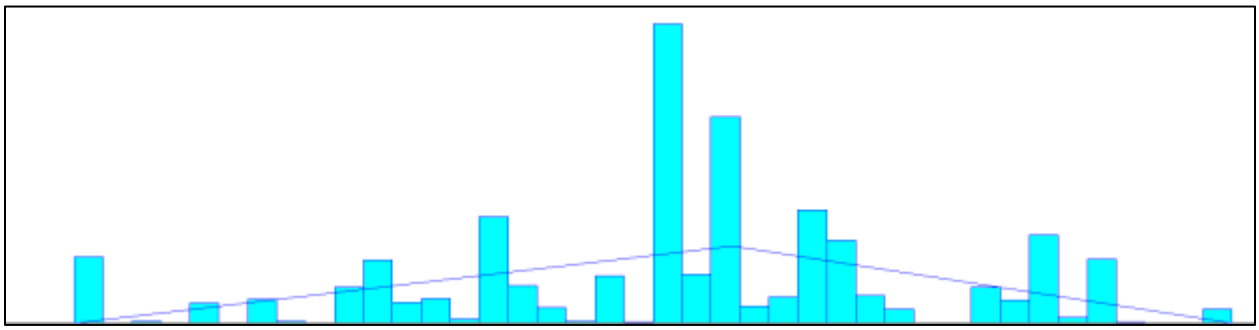
Length-Liner



Distribution Summary	
Distribution:	Lognormal
Expression:	67 + LOGN(157, 58.4)
Square Error:	0.015574
Chi Square Test	
Number of intervals	= 34
Degrees of freedom	= 31
Test Statistic	= 2.52e+03
Corresponding p-value	< 0.005
Kolmogorov-Smirnov Test	
Test Statistic	= 0.0792
Corresponding p-value	< 0.01

Data Summary	
Number of Data Points	= 4992
Min Data Value	= 67
Max Data Value	= 366
Sample Mean	= 222
Sample Std Dev	= 47.1
Histogram Summary	
Histogram Range	= 67 to 366
Number of Intervals	= 40

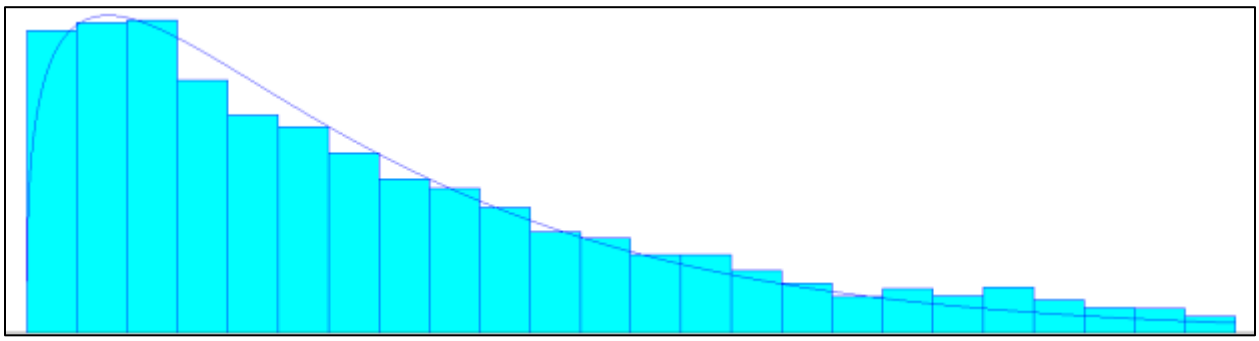
Length-Feeder



Distribution Summary	
Distribution:	Triangular
Expression:	TRIA(50, 158, 240)
Square Error:	0.047589
Chi Square Test	
Number of intervals	= 38
Degrees of freedom	= 36
Test Statistic	= 4.03e+03
Corresponding p-value	< 0.005
Kolmogorov-Smirnov Test	
Test Statistic	= 0.153
Corresponding p-value	< 0.01

Data Summary	
Number of Data Points	= 2573
Min Data Value	= 50
Max Data Value	= 240
Sample Mean	= 149
Sample Std Dev	= 40.5
Histogram Summary	
Histogram Range	= 50 to 240
Number of Intervals	= 40

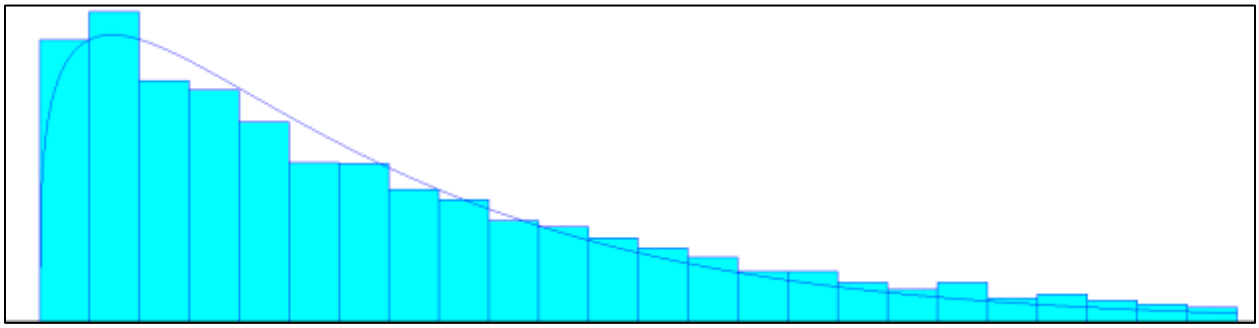
Arrival-Liner



```
        Distribution Summary

Distribution:    Gamma
Expression:      -0.5 + GAMM(5.26, 1.31)
Square Error:    0.000736

Chi Square Test
  Number of intervals    = 24
  Degrees of freedom     = 21
  Test Statistic         = 140
  Corresponding p-value < 0.005
```

```
        Data Summary

Number of Data Points    = 4769
Min Data Value           = 0
Max Data Value           = 23
Sample Mean              = 6.37
Sample Std Dev           = 5.62

        Histogram Summary

Histogram Range          = -0.5 to 23.5
Number of Intervals      = 24
```

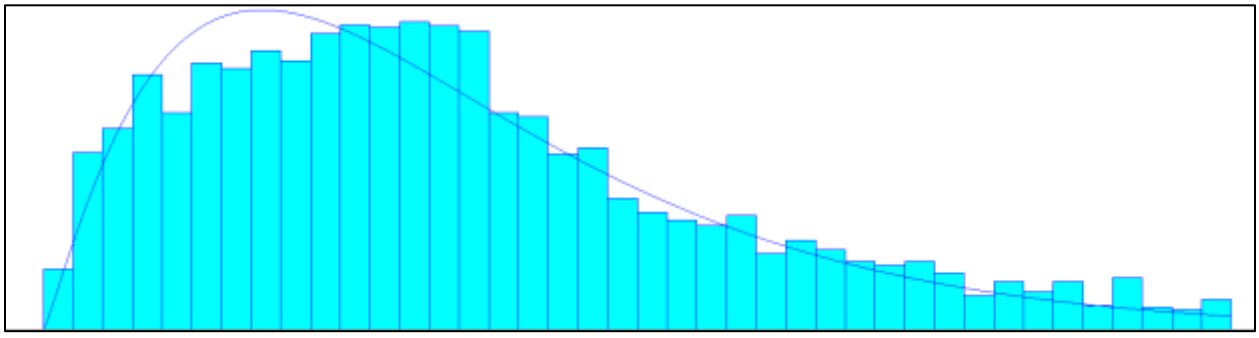
Arrival–Feeder



```
        Distribution Summary

Distribution:    Gamma
Expression:      -0.5 + GAMM(5.27, 1.27)
Square Error:    0.000947

Chi Square Test
  Number of intervals    = 23
  Degrees of freedom     = 20
  Test Statistic         = 69.4
  Corresponding p-value < 0.005
```

```
        Data Summary

Number of Data Points    = 2485
Min Data Value           = 0
Max Data Value           = 23
Sample Mean              = 6.21
Sample Std Dev           = 5.56

        Histogram Summary

Histogram Range          = -0.5 to 23.5
Number of Intervals      = 24
```

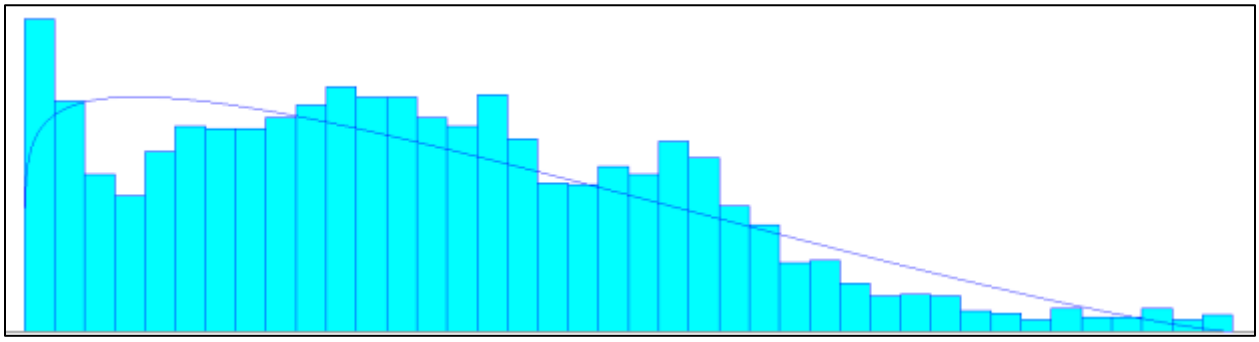
TEU–Liner



```
       Distribution Summary

Distribution:   Gamma
Expression:     5 + GAMM(794, 2.14)
Square Error:   0.000859

Chi Square Test
  Number of intervals   = 40
  Degrees of freedom    = 37
  Test Statistic        = 177
  Corresponding p-value < 0.005

Kolmogorov-Smirnov Test
  Test Statistic        = 0.0316
  Corresponding p-value < 0.01
```

```
        Data Summary

Number of Data Points   = 4699
Min Data Value          = 5
Max Data Value          = 4.94e+03
Sample Mean             = 1.71e+03
Sample Std Dev          = 1.08e+03

       Histogram Summary

Histogram Range         = 5 to 4.94e+03
Number of Intervals     = 40
```

TEU-Feeder



```
       Distribution Summary

Distribution:   Beta
Expression:     2 + 3.01e+03 * BETA(1.13, 2.28)
Square Error:   0.002152

Chi Square Test
  Number of intervals   = 37
  Degrees of freedom    = 34
  Test Statistic        = 207
  Corresponding p-value < 0.005

Kolmogorov-Smirnov Test
  Test Statistic        = 0.0544
  Corresponding p-value < 0.01
```

```
        Data Summary

Number of Data Points   = 2530
Min Data Value          = 2
Max Data Value          = 3.01e+03
Sample Mean             = 997
Sample Std Dev          = 642

       Histogram Summary

Histogram Range         = 2 to 3.01e+03
Number of Intervals     = 40
```

از آنجایی که داده‌ها iid نبودند، نمیتوان با اطمینان کامل توزیع مناسبی را برای داده‌ها در نظر گرفت. این امر باعث شده که آماره ها نیز نتوانند پذیرش فرض صفر مبنی بر فیت بودن داده‌ها را به خوبی تایید کنند. تفکیک داده‌ها بر حسب نوع کشتی نتایج بهتری را حاصل کرد ولی لزوما باعث بهتر شدن آماره آزمون‌ها نشد. در صورتی که نتوان به نتایج خوبی از توزیع داده‌ها رسید بهتر است از توزیع تجربی برای توصیف داده‌های ورودی بهره بگیریم. این کار نیز به کمک نرم‌افزار arena قابل انجام است.