

# PROJECT PART I-DATA COLLECTION AND DESCRIPTIVE STATISTICS

Submitted by

**ARUN JIJU JOSEPH**

**(1002140653)**

in partial fulfilment for the award of the degree of

**M.S. IN DATA SCIENCE**




**THE UNIVERSITY OF TEXAS AT ARLINGTON**

**DASC 5302 INTRODUCTION TO PROBABILITY AND STATISTICS**

**INSTRUCTOR: DR. OBIAGELI LAWRENTIA NGWU**

**SEPTEMBER 2023**

Honor Code: I Arun Jiju Joseph did not give or receive any assistance on this project,  
and the report submitted is wholly my own. 

## Contents

Introduction .....	3
Data Collection.....	4
Descriptive Statistical Analysis.....	5
Data set 1:.....	5
Data set 2:.....	8
Conclusion.....	11
Appendix A.....	12
Data set 1.....	12
Appendix B.....	15
Data set 2.....	15
References .....	18

## **Introduction**

Examining real-world data is the major goal of the research. I looked into and studied two separate sets of actual data for the purpose of this study. Dataset 1 includes the total bill amount for a sample of 100 Starbucks customers. Dataset 2 examined the time difference between two students who entered the library through the main door and was collected at the University of Texas at Arlington Central Library. Data is statistically compiled and visualised to help better understand the trend of the population from the sample.

## **Data Collection**

The bill amount of 100 customers at Starbucks restaurant in The University of Texas at Arlington campus was taken for Data set 1 on September 15, 2023. I noted the customers total bill amount since I had been working in the register section of the restaurant and explained that the information was for a project. The 100 consecutive customers represent a various group of people, including students, faculty and other staffs. By correctly documenting the data collecting process, the data is carefully collected. The bill amount was noted along with the name and gender. The document had the following columns: "Name" (noted while the customers gave their order), "Gender" (collected by asking about how each respondent identified themselves), and "Bill amount" (gathered from order machine after taking their orders). After the data had been effectively gathered, descriptive statistics were used to examine the data for any patterns. This comprises, among other things, calculating the cumulative and relative frequency histograms, assembling the data into a table, and analysing the sample mean and standard deviation.

The University of Texas at Arlington library was the source of the information for Dataset 2 on September 18, 2023, between 11:30 AM and 12 PM. The university's library is one of the busiest places because all students use it for both academic and recreational purposes. The library has got six floors or levels which are used for studying and other purposes. This analysis makes use of information acquired on students using the library. The beginning of the event, which is referred to as the start time, is when the student arrives. About 101 consecutive occasions are recorded every time a student enters the library. A timer on my personal phone and a clock app were used to take the samples. A new gap between two pupils entering the library was represented by each lap. Once the data collection was complete, the time intervals were exported to an Excel file. The following columns were included in the Excel file that was created: "Time" (various time intervals), "Time Difference in Seconds" (time difference discovered by subtracting the first interval from the second interval), and "Seconds" (Seconds). After the data had been effectively gathered, descriptive statistics were used to examine the data for any patterns.

## Descriptive Statistical Analysis

Descriptive analysis is essential to draw conclusions from the data and exhibiting them that are meaningful. It also helps to draw attention to any possible links between the variables. Using the datasets gathered and processed in Excel to perform descriptive statistics using following excel function: **QUARTILE ()**, **AVERAGE ()**, **MEDIAN ()**, **STDEV.S ()**, **VAR.S ()**, **MODE ()**.

As we move on to the Relative Frequency Table, we can see that the data was split up into a number of classes, and the online descriptive statistics calculator was used to figure out how many records there in each class. The ratio of each class relative frequency is determined by dividing its count by the total number of records. The class intervals and relative frequency columns from tables were used to construct a relative frequency histogram. The same class intervals and cumulative relative frequency columns were also used to create a cumulative relative frequency histogram. Data that is provided as histograms are plotted using Excel's built-in histogram function.

### DATASET -1

Statistics Value Units	Values	Units
Sample Mean	6.65	\$
Sample Median	6.42	\$
Sample Mode	6.12	\$
Sample Variance	10.51	\$
Sample Standard Deviation	3.24	\$
Sample Range	17.65	\$
Quartile: Q1	3.99	\$
Quartile: Q2	6.42	\$
Quartile: Q3	8.77	\$

*Table 1.1 Descriptive Statistics for Evaluating the bill amount of 100 customers within UTA Starbucks restaurant*

Table 1.1 demonstrates the sample data's measurements of variability and central tendency. These tabular data are used to create the relative frequency and cumulative relative frequency tables, as well as subsequent visualisations of the same.

Tabular Summary of the Dataset 1			
Class Interval	Count/Frequency	Relative Frequency	Cumulative Relative Frequency
1 – 4	25	0.25	0.25
4 – 7	34	0.34	0.59
7 – 10	27	0.27	0.86
10 – 13	10	0.1	0.96

13-16	3	0.03	0.99
16 - 19	1	0.01	1
<b>Sum</b>	<b>100</b>	<b>1.00</b>	<b>1.00</b>

Table 1.2 Tabular Summary of the Data set 1 combined into different intervals.

Table 1.2 displays the frequency count of customers for each interval. To get the relative frequency, divide each frequency count by the total frequency. To calculate the cumulative frequency, the relative frequencies of the subsequent periods are continuously added.

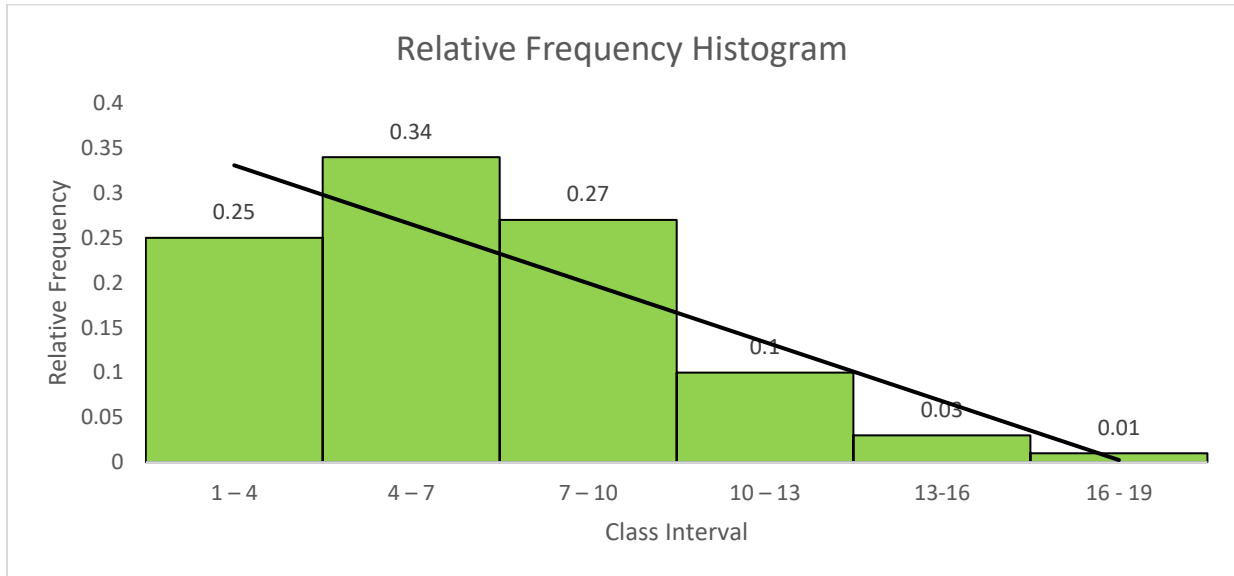


Figure 1.1 Relative Frequency Histogram of Data set 1

The relative frequency histogram in Figure 1.1 reveals the proportion of customers who fall into each group. The majority of customers bills in this particular instance range from \$4 to \$7, which accounts for the 6.42 and 6.65 values for the median and mean of the data. Only 1% of the sample data, or relatively a single customer, have bill between 16 and 19 dollars.

To be regarded as normal, the distribution must be symmetrical with the mean centre. As a result, **the distribution is not symmetric or a normal distribution, it is skewed**, that is, the  $\text{mode} < \text{median} < \text{mean}$  which has a tendency for outliers to right. So, it is **right skewed** which will have the mean to the **right** of the median.

Figure 1.2 below displays the cumulative frequency histogram for bill amount. In cumulative frequency histograms, the number of values in each interval as well as all lower intervals are shown for each bar height.

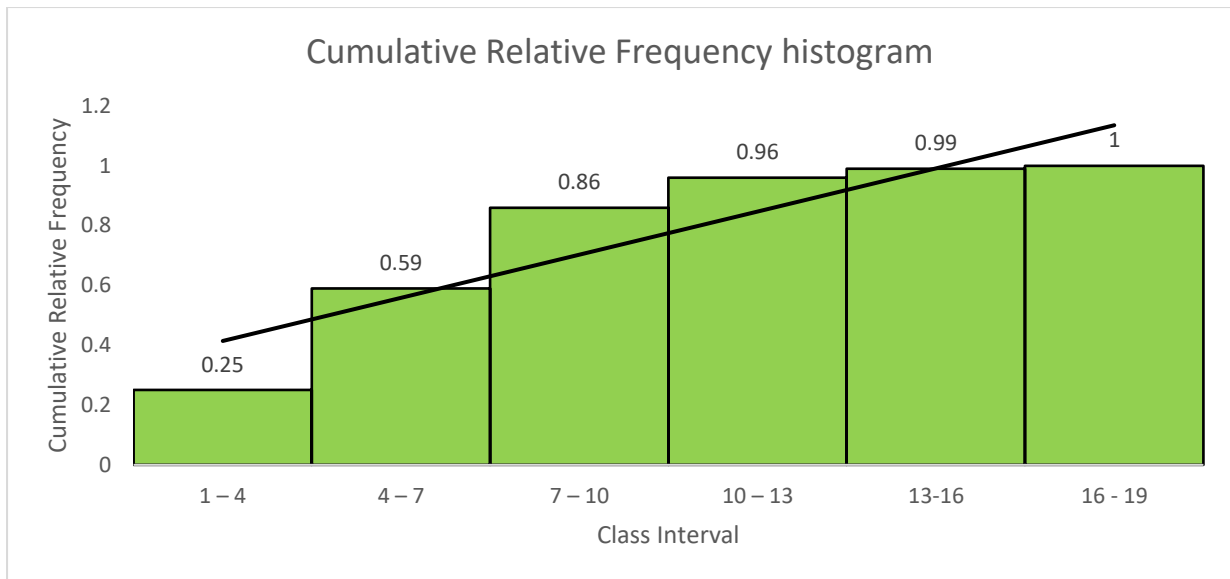


Figure 1.2 Cumulative Relative Frequency histogram for Data set 1

In Figure 1.3, a box and whisker plot are made to more clearly illustrate the whole sample of data. An overview of a set of data is delivered in five numbers. The first quartile is 3.995, the median bill amount is 6.415, the third quartile is 8.775, and the maximum is 18.95. The minimum bill amount for customers is 1.3. Box charts are useful because they reveal outliers in a data collection. An outlier is an observation that differs numerically from the rest of the data. 18.95 dollars being observed in a bill is exceptional, hence it is noted as an anomaly.

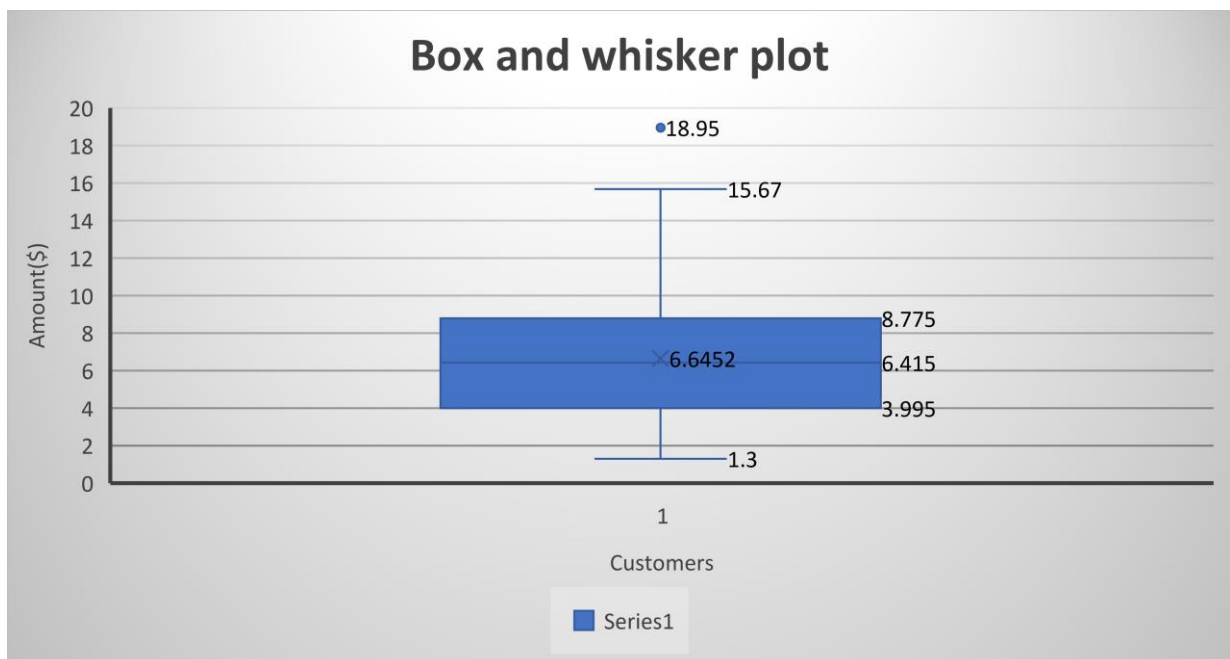


Figure 1.3 Box and whisker plot for Data set 1

## **DATASET 2**

Statistics Value Units	Values	Units
Sample Mean	7.18	Seconds
Sample Median	6	Seconds
Sample Mode	5	Seconds
Sample Variance	26.45	Seconds
Sample Standard Deviation	5.14	Seconds
Sample Range	31	Seconds
Quartile: Q1	4	Seconds
Quartile: Q2	6	Seconds
Quartile: Q3	9	Seconds

*Table 2.1 Descriptive Statistics Exploring the time interval between students entering the UTA Central Library*

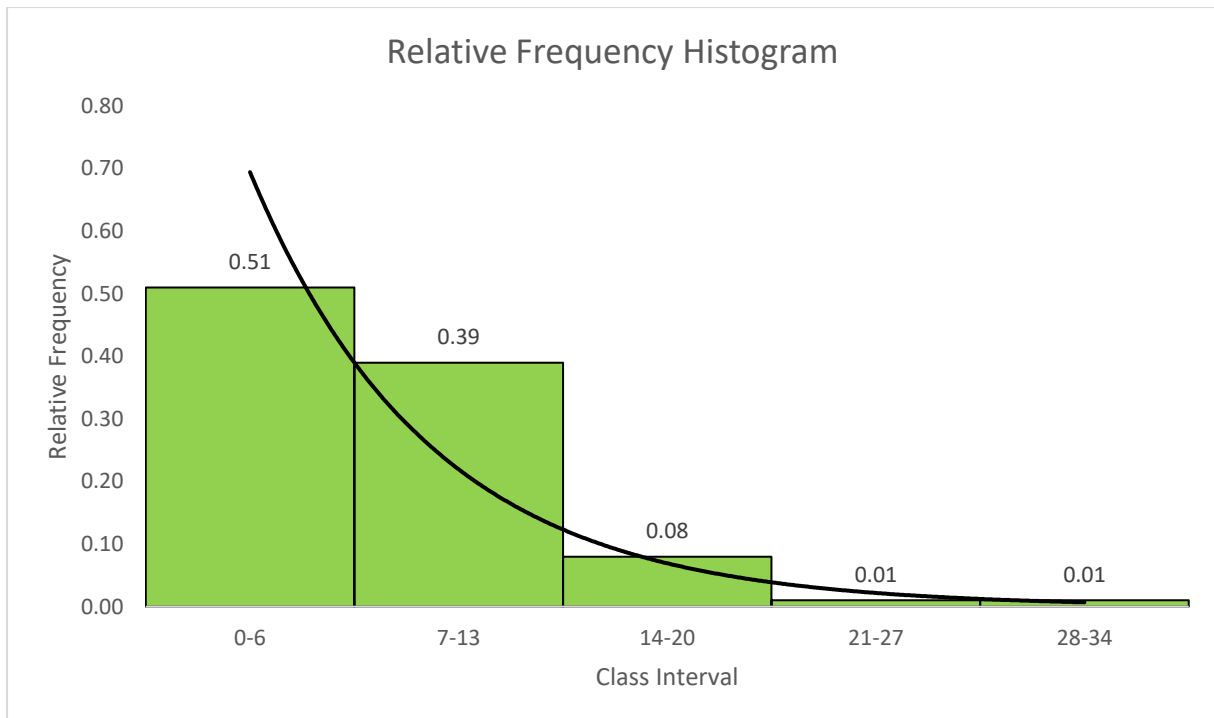
Table 2.1 displays the descriptive statistics values for the intervals between students entering the library. Here, a student enters the library on average in 7.18 seconds. While 51% of students enter in the first six seconds.

Tabular Summary of the Dataset 2			
Class Interval	Count/Frequency	Relative Frequency	Cumulative Relative Frequency
0-6	51	0.51	0.51
7-13	39	0.39	0.90
14-20	8	0.08	0.98
21-27	1	0.01	0.99
28-34	1	0.01	1
<b>Sum</b>	<b>100</b>	<b>1.00</b>	<b>1.00</b>

*Table 2.2 Tabular Summary of the Data set 2 combined into different intervals*

The data from Data Set 2 were divided into five classes in Table 2.2 for a tabular overview of the data. The most students enter the library between 0 and 6 seconds, compared to all other intervals combined.





*Figure 2.1 Relative Frequency Histogram of the Dataset 2*

Figure 2.1 demonstrates that as the time interval gets longer, a relative decrease in the number of students using the library is observed. It is expected that the library is crowded given that students enter there on average in 7.18 seconds. According to our sample data, the relative histogram shows that 90% of students come in the library in less than 14 seconds after the first student arrives. Only 2% of the students take more than 20 seconds to enter.

The sample Data set 2 is used to determine the students' arrival time at the library. In this instance, it has an **exponential distribution** for the chosen intervals. Less number of students visit the library as time goes on. The majority of pupils come with a gap that is shorter, usually between 0 and 6 seconds.

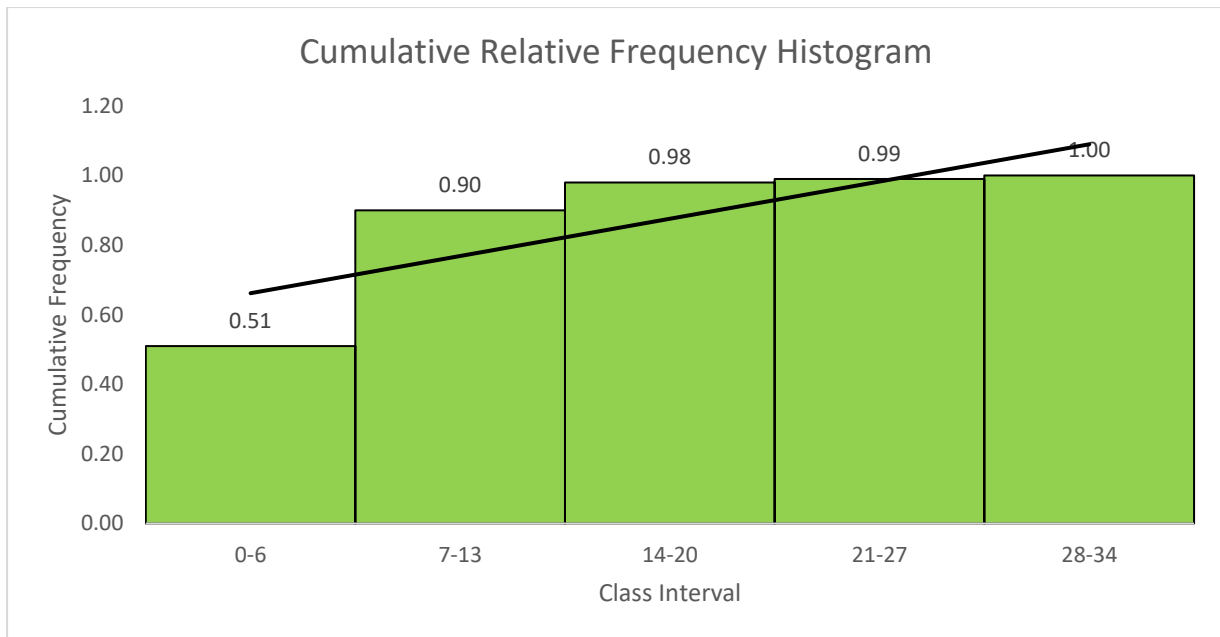


Figure 2.2 Cumulative Relative Frequency Histogram of Data set 2

All students in our sample set enters the library between 0 and 34 seconds following another student, as shown in cumulative relative frequency histogram Figure 2.2.

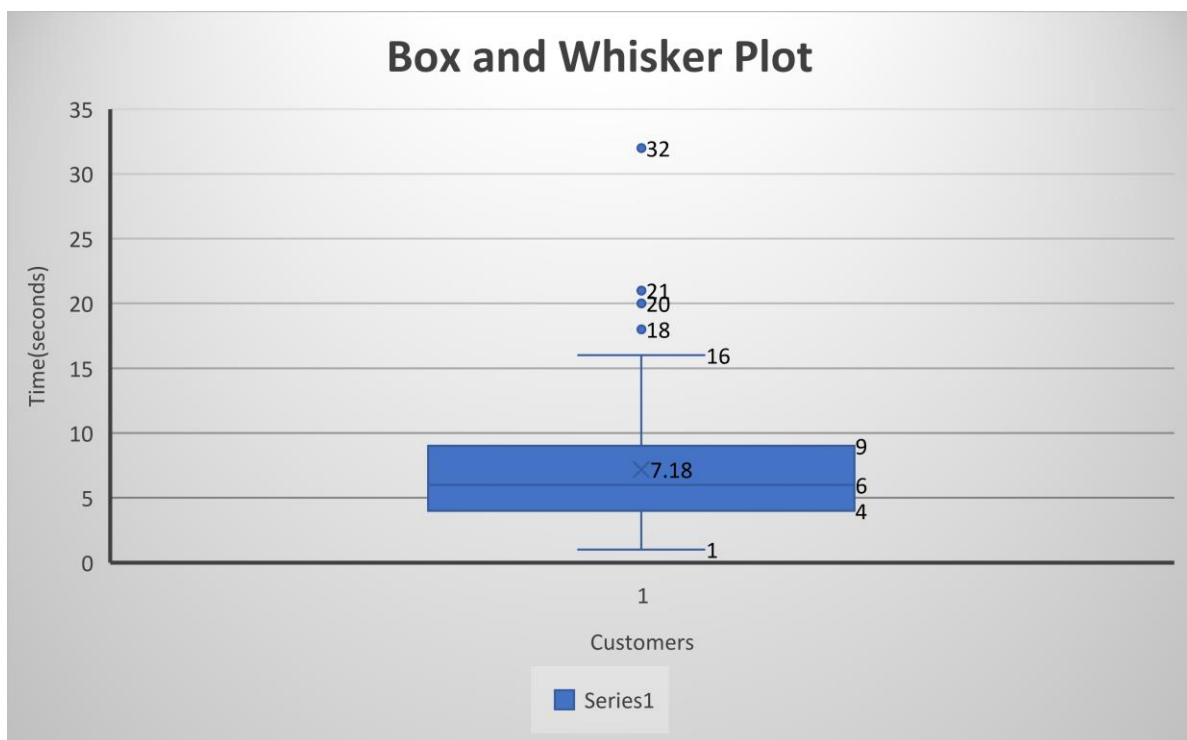


Figure 3.3 Box and Whisker Plot of Data set 2

Outliers can be identified using Data Set 2's box and whisker plot seen in figure 3.3. The sample data outliers in this case are 18, 20, 21, and 32.

## Conclusion

Descriptive analysis uses statistical methods to interpret insights from data by describing it because it might be difficult to extract relevant information from raw data. For data visualisation and data presentation to be effective, descriptive statistics are essential. It is feasible to emphasise any trends in the dataset and ignore outliers by deleting less significant data. I conducted a statistical study on two datasets, starting with the first one that looked at the total bill for a sample of 100 Starbucks customers on September 15, 2023. This analysis's objective was to discover any patterns in the total amount of the 100 consecutive customers' bills. Data Set 2 was collected on September 18, 2023, between 11:30 AM and 12 PM from the library. I obtained information about the student's library arrival timings. The objective of this inquiry is to identify patterns in students' interest in libraries.

Following descriptive statistical analysis on Data Set 1, it was found that the average bill size is 6.65. Many customers bills, as shown in Table 1, ranged from \$4 to \$7. Table 1.2 shows that 86% of customers have bills that are less than \$10. Additionally, Figure 1.1 demonstrates that only one customer, or 1% of the sample data, paid the bill amount between the 16 and 19. These conclusions are based on the fact that the majority of the data was gathered from university students, many of whom had a preference for specific Refreshers and Frappuccino drinks, which had a maximum price of \$8. As a result, the given distribution is a skewed **distribution, that is, it is not a symmetric or a normal distribution**, it has a tendency for outliers to right. The mean, median, and mode must all be equal for there to be a normal distribution; in this case, they are 6.65, 6.42, and 6.12, respectively.

In Data set 2, Descriptive statistics reveal that it takes a student, on average, 7.18 seconds to enter the library. The first 6 seconds show 51% of the students enter. Table 2.1 shows that compared to all other times, the most students enter the library between 0 and 6 seconds. Based on our sample data, the relative histogram shows that 90% of students enter the library within 13 seconds of the first student. The mean student arrival time was 7.18 seconds, so the median student arrival time was 6. This is less time than the mean. It so satisfies a necessary condition for an **exponential distribution**. For the selected intervals, an exponential distribution is used. The number of pupils accessing the library is decreasing with time. Many students arrive with shorter gaps, typically between 0 and 6 seconds.

## **Appendix A**

### **Data set 1**

The raw data from Data Set 1 that was collected with different values and used for the descriptive statistical analysis are shown in the table below.

<b>No</b>	<b>NAME</b>	<b>Gender</b>	<b>BILL AMOUNT (\$)</b>
1	Evan	Male	6.12
2	John	Male	3.54
3	Jose	Male	7.01
4	Sam	Male	12.09
5	Kate	Female	6.12
6	Marie	Female	9.01
7	Shaun	Male	10.33
8	Jr	Male	3.45
9	Erik	Male	2.07
10	Erica	Female	8.76
11	Briana	Female	11.43
12	Lewis	Male	10.87
13	Sai	Male	6.78
14	Ram	Male	5.67
15	Jessica	Female	6.12
16	Evana	Female	1.3
17	EKR	Male	2.41
18	Angel	Female	11.32
19	Fathima	Female	6.12
20	Nikita	Female	1.3
21	Laya	Female	2.36
22	Dj	Male	4.32
23	Lillian	Female	8.76
24	Lilly	Female	14.54
25	Sona	Female	3.32
26	Danny	Male	2.68
27	Yesha	Female	7.48
28	Gabriel	Male	5.53
29	Derrick	Male	3.01
30	John	Male	2.98
31	Isaac	Male	3.78
32	Ben	Male	4.66
33	Nora	Female	7.91
34	Lesley	Female	3.21
35	Erin	Female	4.34
36	Mathew	Male	5.32
37	Anna	Female	4.88
38	Ashley	Female	2.56
39	Jenna	Female	3.78
40	Lily	Female	6.7
41	Jason	Male	8.99
42	Juan	Male	5.45

43	Son	Male	6.75
44	Mellany	Female	3.45
45	Jenna	Female	6.32
46	Nora	Female	3.32
47	Sara	Female	4.19
48	Monica	Female	8.44
49	Michael	Male	7.31
50	Jake	Male	9.48
51	Jetty	Male	11.43
52	Ken	Male	6.51
53	Cassey	Female	3.32
54	Tj	Male	15.67
55	Edwin	Male	8.66
56	Chang	Male	10.41
57	Joanna	Female	2.58
58	Alen	Male	6.67
59	Dk	Male	7.66
60	Lee	Male	5.41
61	Mona	Female	6.14
62	Kendra	Female	8.16
63	Alisa	Male	9.88
64	Sona	Female	10.19
65	Katelyn	Female	8.94
66	Robert	Male	7.89
67	Charles	Male	18.95
68	Elizabeth	Female	10.42
69	Suffiya	Female	3.45
70	Bhavya	Female	6.12
71	Anita	Female	7.17
72	Sharon	Female	9.01
73	Edward	Male	8.76
74	Amy	Female	9.12
75	Maria	Female	10.12
76	Diane	Female	2.67
77	Jacqueline	Female	8.9
78	Sean	Male	5.11
79	Kathryn	Female	6.11
80	Lawrence	Male	2.91
81	Russell	Male	6.88
82	Doris	Female	3.51
83	Scott	Male	4.31
84	Stephen	Male	5.51
85	Gregory	Male	6.71
86	Omar	Male	8.78
87	Janet	Female	7.71
88	Jerry	Male	1.89
89	Kelly	Female	7.93
90	Peter	Male	7.51
91	Gerald	Male	4.87

92	Jordan	Male	13.45
93	Mary	Female	8.97
94	Claire	Female	4.54
95	Rohan	Male	6.88
96	Shine	Male	4.64
97	Stephen	Male	6.78
98	Kevin	Male	8.93
99	Ben	Male	4.81
100	Joseph	Male	3.93

## **Appendix B**

### **Data set 2**

The raw data from Data Set 2 that was collected with different values and used for the descriptive statistical analysis are shown in the table below.

<b>No</b>	<b>Time</b>	<b>Time Difference (Seconds)</b>	<b>Seconds</b>
1	11:35:29	0	0
2	11:35:36	00:00:07	7
3	11:36:08	00:00:32	32
4	11:36:20	00:00:12	12
5	11:36:25	00:00:05	5
6	11:36:34	00:00:09	9
7	11:36:50	00:00:16	16
8	11:37:00	00:00:10	10
9	11:37:21	00:00:21	21
10	11:37:32	00:00:11	11
11	11:37:48	00:00:16	16
12	11:37:53	00:00:05	5
13	11:38:00	00:00:07	7
14	11:38:05	00:00:05	5
15	11:38:09	00:00:04	4
16	11:38:16	00:00:07	7
17	11:38:21	00:00:05	5
18	11:38:24	00:00:03	3
19	11:38:30	00:00:06	6
20	11:38:50	00:00:20	20
21	11:39:05	00:00:15	15
22	11:39:07	00:00:02	2
23	11:39:15	00:00:08	8
24	11:39:25	00:00:10	10
25	11:39:28	00:00:03	3
26	11:39:30	00:00:02	2
27	11:39:37	00:00:07	7
28	11:39:50	00:00:13	13
29	11:39:58	00:00:08	8
30	11:40:03	00:00:05	5
31	11:40:07	00:00:04	4
32	11:40:25	00:00:18	18
33	11:40:32	00:00:07	7
34	11:40:35	00:00:03	3
35	11:40:40	00:00:05	5
36	11:40:48	00:00:08	8
37	11:40:55	00:00:07	7
38	11:41:07	00:00:12	12
39	11:41:25	00:00:18	18
40	11:41:30	00:00:05	5
41	11:41:35	00:00:05	5

42	11:41:40	00:00:05	5
43	11:41:48	00:00:08	8
44	11:41:50	00:00:02	2
45	11:41:57	00:00:07	7
46	11:41:59	00:00:02	2
47	11:42:02	00:00:03	3
48	11:42:10	00:00:08	8
49	11:42:21	00:00:11	11
50	11:42:24	00:00:03	3
51	11:42:29	00:00:05	5
52	11:42:33	00:00:04	4
53	11:42:35	00:00:02	2
54	11:42:39	00:00:04	4
55	11:42:43	00:00:04	4
56	11:42:47	00:00:04	4
57	11:42:48	00:00:01	1
58	11:42:56	00:00:08	8
59	11:42:58	00:00:02	2
60	11:43:00	00:00:02	2
61	11:43:12	00:00:12	12
62	11:43:15	00:00:03	3
63	11:43:20	00:00:05	5
64	11:43:38	00:00:18	18
65	11:43:43	00:00:05	5
66	11:43:47	00:00:04	4
67	11:43:49	00:00:02	2
68	11:43:50	00:00:01	1
69	11:43:53	00:00:03	3
70	11:44:02	00:00:09	9
71	11:44:08	00:00:06	6
72	11:44:09	00:00:01	1
73	11:44:17	00:00:08	8
74	11:44:25	00:00:08	8
75	11:44:28	00:00:03	3
76	11:44:30	00:00:02	2
77	11:44:40	00:00:10	10
78	11:44:44	00:00:04	4
79	11:44:48	00:00:04	4
80	11:44:57	00:00:09	9
81	11:45:05	00:00:08	8
82	11:45:12	00:00:07	7
83	11:45:25	00:00:13	13
84	11:45:30	00:00:05	5
85	11:45:38	00:00:08	8
86	11:45:39	00:00:01	1
87	11:45:41	00:00:02	2
88	11:45:48	00:00:07	7
89	11:45:58	00:00:10	10
90	11:46:05	00:00:07	7
91	11:46:20	00:00:15	15



92	11:46:28	00:00:08	8
93	11:46:33	00:00:05	5
94	11:46:37	00:00:04	4
95	11:46:45	00:00:08	8
96	11:46:52	00:00:07	7
97	11:46:55	00:00:03	3
98	11:47:06	00:00:11	11
99	11:47:18	00:00:12	12
100	11:47:23	00:00:05	5
101	11:47:27	00:00:04	4

## **References**

1) <https://www.simplilearn.com/what-is-descriptive-statisticsarticle#:~:text=Descriptive%20statistics%20refers%20to%20a%20set%20of%20methods%20used%20to,help%20identify%20patterns%20and%20relationships.>

2) <https://www.hackmath.net/en/calculator/frequency-table>

3) <https://www.geeksforgeeks.org/descriptive-statistic/>

4) <https://goodcalculators.com/grouped-frequency-distribution-calculator/>