Hands on Assignment #1

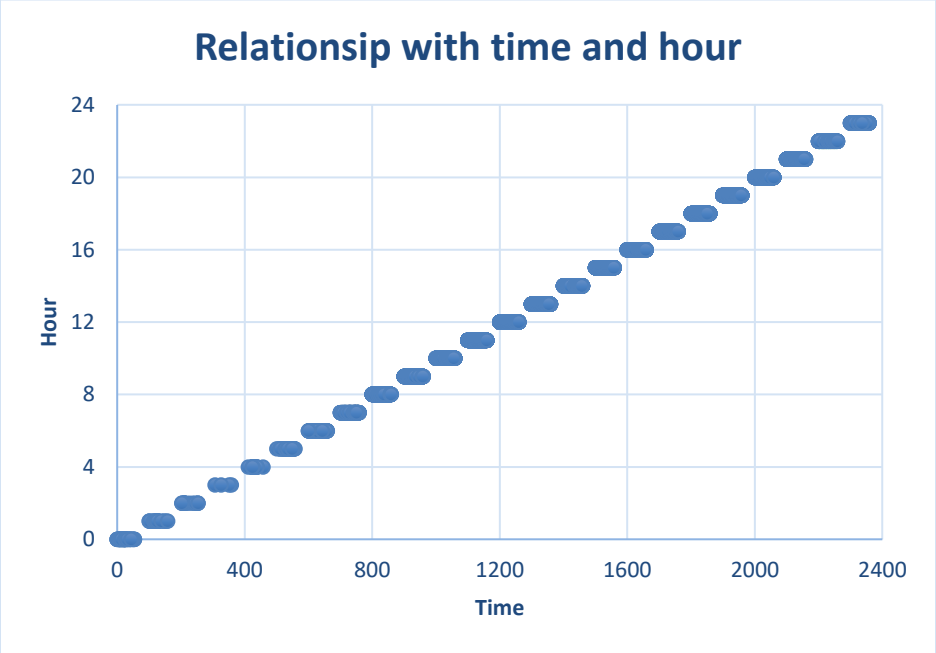
The “ATM Transactions” Excel Workbook contains a data set of Credit union transactions for 30 days at a drive-up ATM and two walk-up ATMs on-campus in student union buildings at two different university campuses. Create a comprehensive data analysis report that addresses the following questions. Please make sure that all of your visualizations have proper labels, titles, axis titles, and that they follow general guidelines of creating good data visualization.

1) Describe the data, including how many total observations there are.
   Answer: The data set contains the credit union transactions at an ATM for 3 different locations for one month. It has 14,913 observations with different types of transactions like deposits, withdrawals, transfer and advance.
   We have 8 types of data with the following observations as shown below:
   1. Time
   2. Type
   3. Date
   4. Day Code
   5. Weekday
   6. Location
   7. Hour
   8. Amount

2) What do you think the field “time” means? How would you interpret it?
   Answer: Time field shows the time of transaction in atm in “hour: minute” format. In “hour” field, we can see that if hour start with 6 then time also start with 6 or vice versa which conclude that time is shown in “hour:minute”. For example, if time is mentioned as 625, then it means the time at which transaction happened in atm is 6:25 am. Another example; if time shows 1730 then it means time is 17:30 or 5:30pm.

3) Create a visual (some sort of graph) with the variable time to test your hypothesis.

Answer: Scatter plot for time field is shown below:



4) Include tables that count the number of observations by:

i) Weekday

ii) Location

iii) Type of transaction

Answer:

- For Weekday, table is shown below:

Weekday	No. of observations by weekday
Sun	1117
Mon	1892
Tue	2522
Wed	2720
Thu	2377
Fri	2801
Sat	1484
Grand Total	**14913**

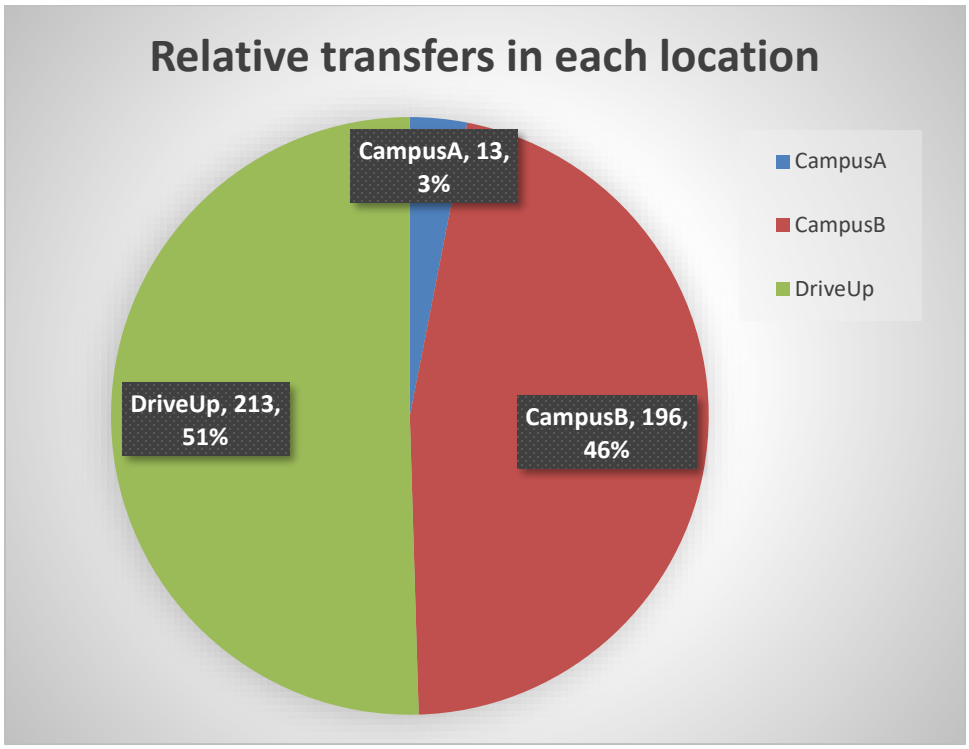
- For Location, table is shown below:

Location	Count of location
CampusA	658
CampusB	7285
DriveUp	6970
Grand Total	**14913**

- For type of transaction, table is shown below:

Types of Transaction	Count of each type of transaction
Advance	114
Deposit	3662
Transfer	422
Withdrawal	10715
Grand Total	**14913**

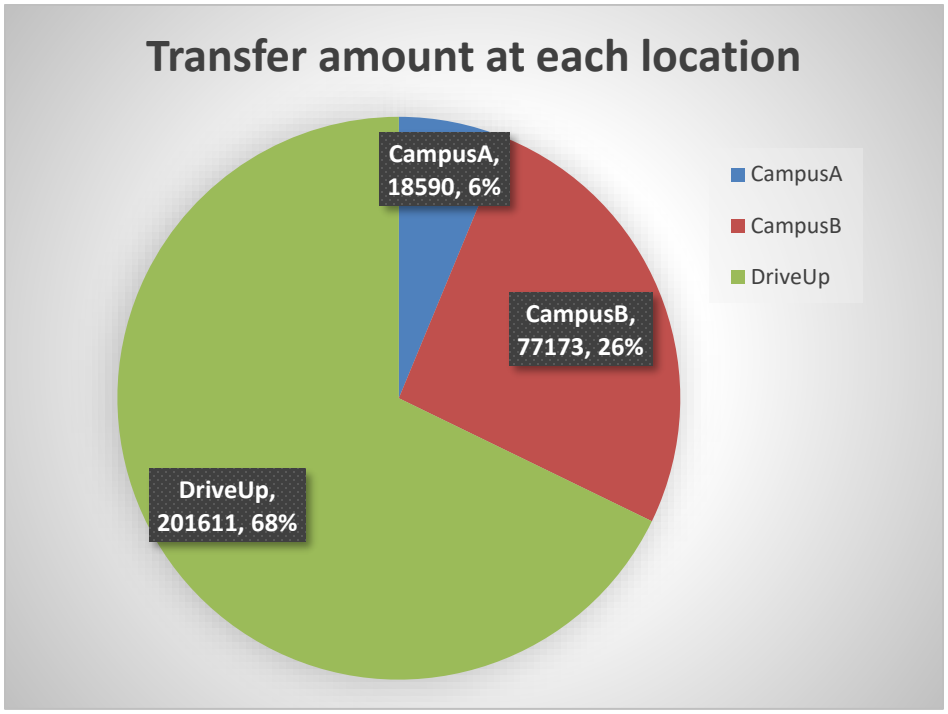
5) Create a pie chart to illustrate the relative **number** of transfers in each location.
Pie Chart 1



Location	No. of Transfer
CampusA	13
CampusB	196
DriveUp	213
Grand Total	**422**

6) Create a pie chart to illustrate the relative total transfer amount at each location.

Pie chart 2



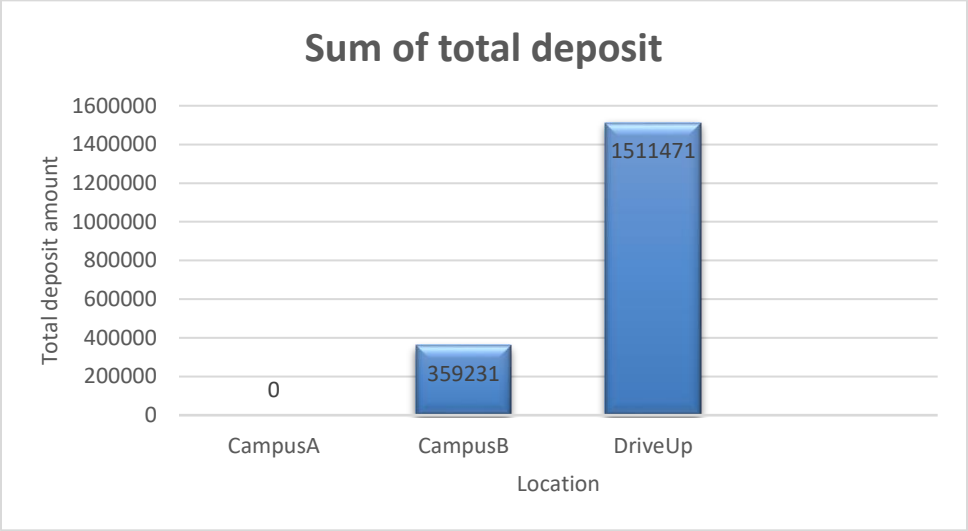
Location	Total Transfer amount
CampusA	18590
CampusB	77173
DriveUp	201611
Grand Total	**297374**

7) Discuss the differences in the two pie charts. What do they tell you?

Answer: Pie chart1 describes the relative number of transfers while the Pie chart 2 shows the amount transferred at three locations. Pie chart 1 illustrates that number of transactions were least in CampusA (3%) and most in Driveup (51%). Same pattern was followed in Pie chart 2 which says that DriveUp has highest transferred amount (68%) compared to CampusA and CampusB. Though, difference in "number of transfers" between DriveUp and CampusB is just 17

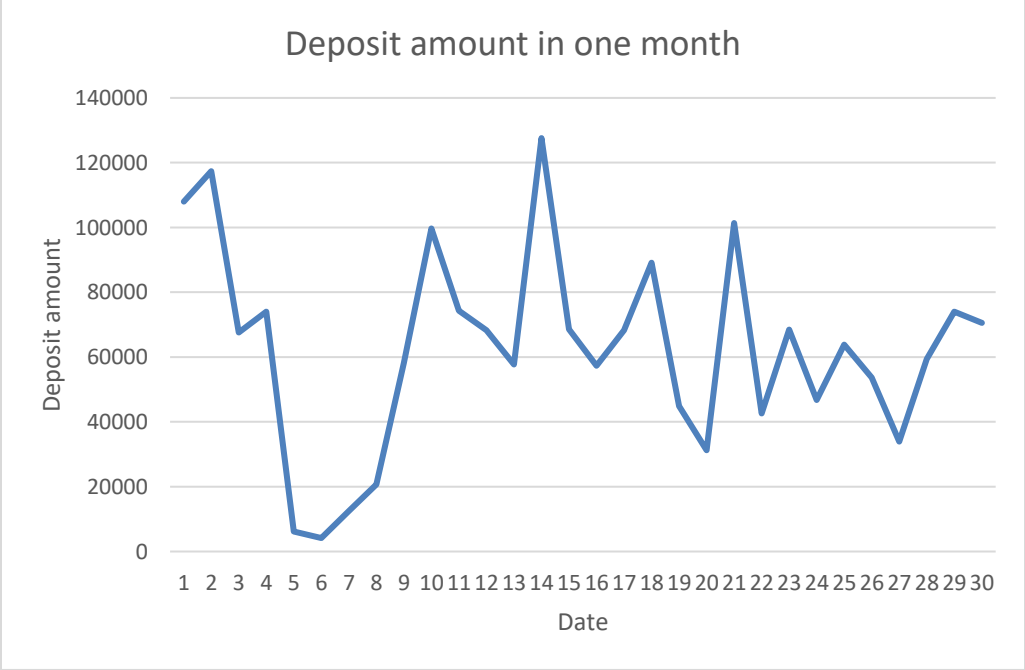
(213-196), the amount transferred in CampusB is far lesser than DriveUp which considerably lowering the percentage of amount transfer (just 26%) in pie chart 2.

8) Create a column chart to illustrate the total deposits for the month in each location.



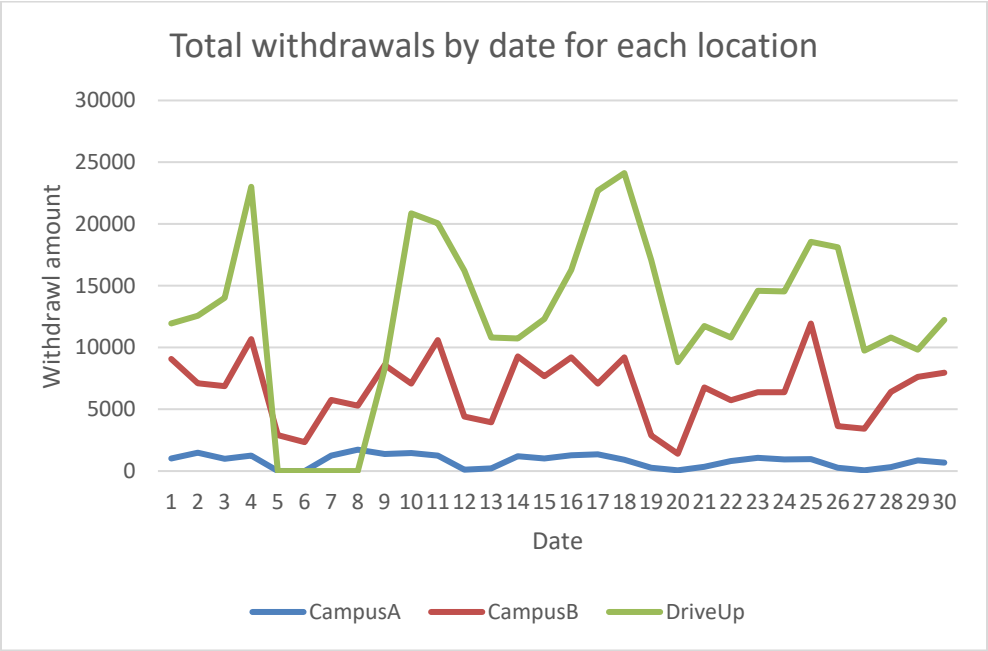
Location	Sum of total deposit
CampusA	0
CampusB	359231
DriveUp	1511471
Grand Total	**1870702**

9) This data contains 30 days of transactions, with the "date" variable indicating the date of the transaction. Create a chart for the time series of deposit totals by date.



Date	Deposit amount
1	108042
2	117415
3	67574
4	73983
5	6129
6	4160
7	12541
8	20719
9	58202
10	99663
11	74326
12	68257
13	57718
14	127590
15	68605
16	57377
17	68337
18	89166
19	44899
20	31258
21	101358
22	42602
23	68499
24	46754
25	63891
26	53670
27	33971
28	59397
29	74000
30	70599
Grand Total	**1870702**

10) Create a chart that illustrates the time series of total withdrawals by date for each location. (There should be 3 lines on this chart, since there are 3 ATM locations.)



Date	CampusA	CampusB	DriveUp	Total amount withdrawn in each day
1	1030	9070	11940	22040
2	1480	7100	12570	21150
3	990	6860	14010	21860
4	1240	10690	23010	34940
5	10	2900	0	2910
6	0	2350	0	2350
7	1250	5750	0	7000
8	1730	5300	0	7030
9	1390	8580	8280	18250
10	1460	7070	20855	29385
11	1260	10600	20060	31920
12	110	4420	16220	20750
13	220	3950	10810	14980
14	1200	9280	10730	21210
15	1020	7670	12310	21000
16	1270	9200	16286	26756
17	1350	7080	22710	31140

18	910	9200	24120	34230
19	270	2890	17090	20250
20	60	1400	8810	10270
21	350	6760	11740	18850
22	800	5720	10810	17330
23	1060	6370	14590	22020
24	940	6370	14547	21857
25	970	11940	18550	31460
26	280	3640	18110	22030
27	50	3430	9740	13220
28	330	6410	10810	17550
29	850	7620	9830	18300
30	690	7950	12240	20880
Grand Total	**24570**	**197570**	**380778**	**602918**

For #11 – 24, consider only Withdrawal transactions.

11) Create a histogram for the amount withdrawn.

Histogram for withdrawn amount. (Total number of outliers were 1037)

To calculate outlier,

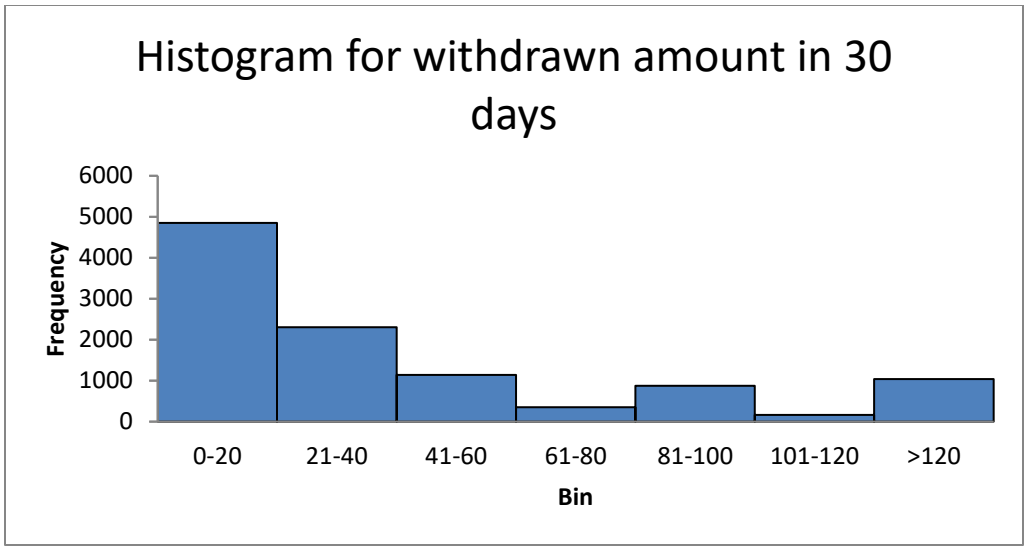
Q1: 20, Q2: 30, Q3: 60

IQR = Q3 – Q1 = 40

Lower Limit = Q1 – IQR*1.5 = -40

Upper Limit =Q3 + IQR*1.5 = 120

Bin	*Frequency*
0-20	4850
21-40	2304
41-60	1140
61-80	350
81-100	873
101-120	161
>120	1037

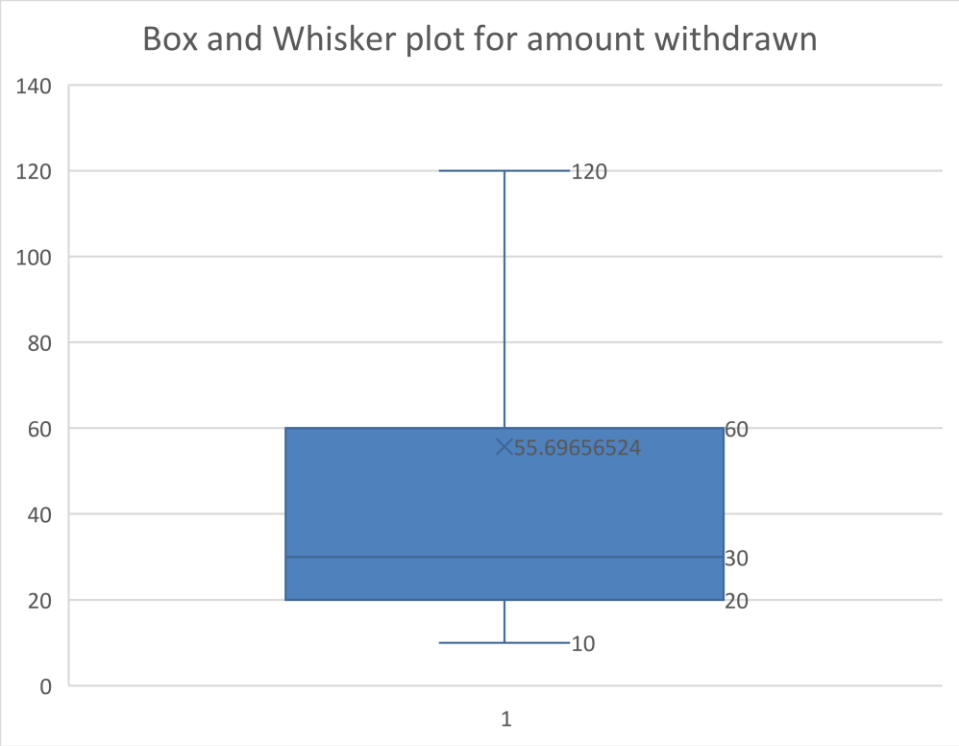


12) What is the average amount withdrawn?
Answer: 56.2686

13) What is the median amount withdrawn?
Answer: 30

14) What is the standard deviation of the amount withdrawn?
Answer: 92.8047

15) Create the five-number summary and draw the box and whiskers plot for the amount withdrawn.
Box and Whisker plot. (Outlier: 6185 was excluded while calculating the quartiles)

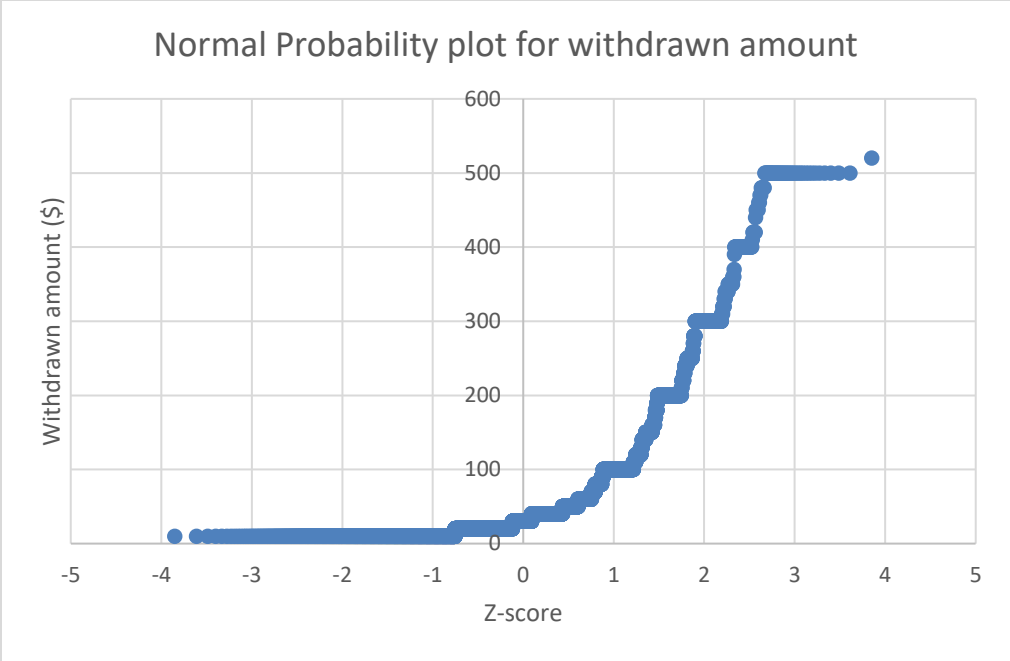


Five number summary:
Min(Q0): 10
Q1: 20
Median(Q2): 30
Q3: 60
Max(Q4): 520
In this case, Max will be 120 as it is upper limit (calculated below)

IQR = Q3 – Q1 = 40
Lower Limit = Q1 – IQR*1.5 = -40 (Or Q0, whichever is bigger)
Upper Limit =Q3 + IQR*1.5 = 120 (Or Q4 whichever is smaller)

16) Describe the distribution and general shape of the amount withdrawn.
Answer: The shape is right skewed where mean(56.2686) is greater than median(30). It has longer tail on right side. Median is close to first quartile. Most of the data is distributed on left side.

17) Create a normal probability plot for the amount withdrawn.



18) Does the distribution look normal? Why or why not?
Answer: The distribution is not normal. It is not symmetric. It is right skewed. Also, it is having right in-wards curve which is going up (as shown in above graph).

19) Are there any outliers? How many?
Answer: Yes. Total number of Outliers are 1037.

20) From this data, create a contingency table for withdrawals over $50 / =$50 /under $50 in one dimension, and location on the other.

Table1:

Location	Number of withdrawn in each location			
	<50	50	>50	Grand total
CampusA	492	31	121	644
CampusB	4524	274	759	5557
DriveUp	2139	338	2037	4514
Grand Total	**7155**	**643**	**2917**	**10715**

21) What is the probability that a withdrawal is >= $50?
Answer: 0.33

22) If a withdrawal is over $50, what is the probability that it originated in the drive-up location?
Answer: 0.698

23) What is the probability that a withdrawal from Campus A is greater than $50?
Answer: 0.1878

24) Are location and size of withdrawal (>50, =50, <50) independent events?
Answer: No

Please check below table for relative probability: (Relative probability was calculated with the help of table1 (in question 20) shown above)

Location	Relative Probabilities			
	<50	50	>50	Grand total
CampusA	0.045917	0.002893	0.011293	0.06010266
CampusB	0.422212	0.025572	0.070835	0.51861876
DriveUp	0.199627	0.031545	0.190107	0.42127858
Grand Total	**0.667755**	**0.060009**	**0.272235**	**1**

P(CampusA and <50)

P(<50)

P(CampusA)

For independent events, P(A and B)=P(A).P(B)
P(CampusA and <50)=0.045917
P(CampusA)= 0.06010266
P(<50)=0.667755
In this case, P(CampusA and <50) is not equal to P(CampusA).P(<50)
Similarly, for other cases, location and withdrawal amount are not independent events.

For #25 – 37, consider only Transfer transactions.

25) Create a histogram for the amount transferred.

Answer: Histogram for transferred. (Total number of outlier were 51)

To calculate outlier,

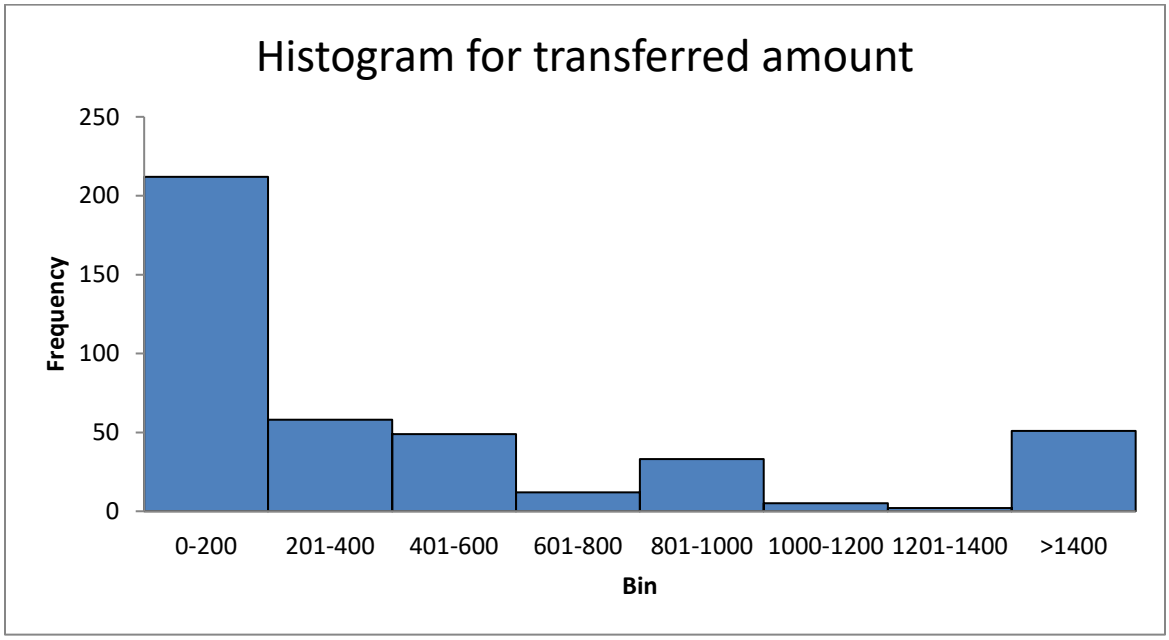
Q1: 75, Q2: 200, Q3: 600

IQR = Q3 – Q1 = 525

Lower Limit = Q1 – IQR*1.5 = -712.5

Upper Limit =Q3 + IQR*1.5 = 1387.5

Bin	*Frequency*
0-200	212
201-400	58
401-600	49
601-800	12
801-1000	33
1000-1200	5
1201-1400	2
>1400	51



26) What is the average amount transferred?

Answer: 704.6777

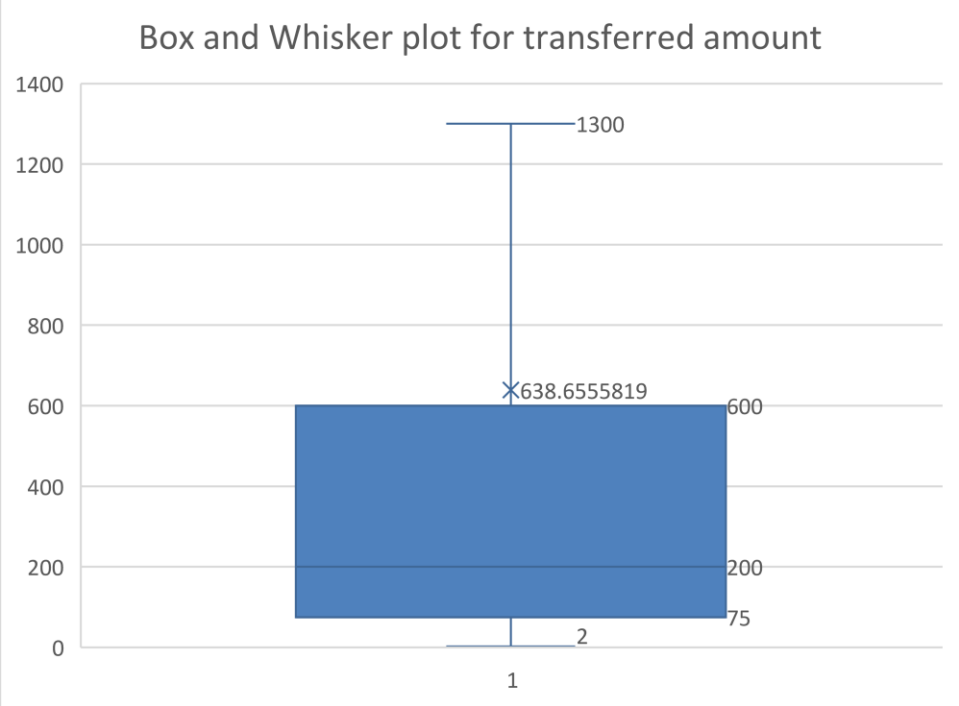
27) What is the median amount transferred?

Answer: 200

28) What is the standard deviation of the amount transferred?
Answer: 1800.209

29) Compare the standard deviations of transfers and withdrawals and talk about the differences in the distributions.
The standard deviation of the withdrawal amount is 92.8047 where mean is 56.2686. It shows standard deviation is closer to mean. On the other hand, for transfer amount, standard deviation (1800.209) is very far from mean(704.6777). The distribution for both is showing right skewed. But transfer amount is heavily skewed towards right compared to withdrawal amount graph. Also, there were lower amount of withdrawal happened compared to transfer amount.

30) Create the five-number summary and draw the box and whiskers plot for the amount transferred.
Box and Whisker plot. (Outlier: 28500 was excluded, while calculating the quartiles)



Five number summary:
Min(Q0): 2
Q1: 75
Median(Q2): 200
Q3: 600
Max(Q4): 10000
In this case, Max will be 1300.

We calculated upper limit as 1387.5. (But, there is no data as 1387.5 in the transaction, so graph is showing upper limit is 1300.)

IQR = Q3 – Q1 = 525
Lower Limit = Q1 – IQR*1.5 = -712.5 (Or Q0, whichever is bigger)
Upper Limit =Q3 + IQR*1.5 = 1387.5 (Or Q4 whichever is smaller)

31) Describe the distribution and general shape of the amount transferred.
Answer: The shape is right skewed where mean (704.6777) is greater than median(200). It has longer tail on right side. Median is close to first quartile. Most of the data is distributed on left side.

32) Does the distribution look normal? Why or why not?
Answer: Distribution does not look normal. It is not symmetric and heavily skewed towards right where mean(704.6777) is greater than median(200) . Most of the data is distributed on left side. Hence, it is not normal distribution.

33) Are there any outliers? How many?
Answer: Yes, Total number of outliers is 51.

34) From this data, create a contingency table for transfers >= $500/under $500 in one dimension and location on the other.
Table2:

Location	Number of transfers in each location		
	<500	>=500	Grand Total
CampusA	5	8	13
CampusB	146	50	196
DriveUp	128	85	213
Grand Total	**279**	**143**	**422**

35) What is the probability that a transfer is >= $500? (i.e. what proportion of all transfers is over $500?)
Answer: 0.338

36) If a transfer is under $500, what is the probability that it originated in the drive-up location?
Answer: 0.458

37) What is the probability that a transfer from Campus B is greater or equal to $500?
Answer: 0.255

38) Are location and size of transfer (over/under $500) independent events?

Answer: No

For independent events, P(A and B)=P(A).P(B)

In this case, P(CampusA and <500)=0.011

P(CampusA)= 0.0308

P(<500)=0.66

In this case, P(CampusA and <500) is not equal to P(CampusA).P(<500)

Similarly, for other cases, location and transfer amount are not independent events.