COMPLETED
SAMPLE OF WORK
DATA SCIENCE
(Explanatory Data Analysis)

NORIE JEANNE PEREIRA

Dataset: data.csv

This process involves generating questions, and investigating them with visualizations.
#EDA is important because it allows you to understand your data, and make unintended discoveries.
#To build an EDA project, keep the following topics in mind:
#1. Formulate relevant questions, and hypotheses
#2. Test those questions with visualizations
#3. Identify trends in the data
#4. Look for relationships between variables
#5. Communicate results with visualizations (scatter plots, histograms, etc.)

Steps:

1. For the preparations lets first import the necessary libraries and load the files needed for our EDA. This includes:
   1. Pandas: wraps NumPy arrays with a series and dataframe object providing lots of convenient methods.
   2. Matplotlib: The most full-featured Python plotting library. Generates static images.

2. Output -read csv

```
 BranchName  Week  DayWeek  Day  Month  Hour Transaction_Type  Units  Amount
0    MyStore     1        3    2      1     9             Card      3   54.00
1    MyStore     1        3    2      1    10             Cash      7  -17.80
2    MyStore     1        3    2      1    10             Card      7   41.99
3    MyStore     1        3    2      1    11             Card     20  412.50
4    MyStore     1        3    2      1    12             Cash      1  -18.00
```

First I removed the $ sign and then converted the string field into numeric, once done, we should have data in float since we are going to perform mathematical operations on this field.

3. Output-Remove unwanted columns say BranchName(One thing more, I see BranchName field unnecessary since we only have data of a single store so let’s remove it!

	Week	**DayWeek**	**Day**	**Month**	**Hour**	**Transaction_Type**	**Units**	**Amount**
0	1	3	2	1	9	Card	3	54.00
1	1	3	2	1	10	Cash	7	-17.80
2	1	3	2	1	10	Card	7	41.99
3	1	3	2	1	11	Card	20	412.50

We already done with cleaning .

4. We need to find the number of records and columns. I try to execute df.shape.
   And found out that there are 4100 total records and 9 columns.
   (4100,9)

5. I need a detailed summary of this data, for that I am going to run df.describe

	Week	DayWeek	Day	Month	Hour	Units	Amount
count	4100.000000	4100.000000	4100.000000	4100.000000	4100.000000	4100.000000	4100.000000
mean	34.017805	4.183902	15.812195	8.231463	12.949024	12.779512	35.237046
std	14.714289	1.967864	8.810817	3.396586	2.631853	17.854968	183.538724
min	1.000000	1.000000	1.000000	1.000000	8.000000	1.000000	-1041.800000
25%	29.000000	3.000000	8.000000	7.000000	11.000000	3.000000	-47.000000
50%	37.000000	4.000000	16.000000	9.000000	13.000000	8.000000	-2.385000
75%	45.000000	6.000000	23.000000	11.000000	15.000000	16.000000	99.512500
max	53.000000	7.000000	31.000000	12.000000	19.000000	274.000000	1487.000000

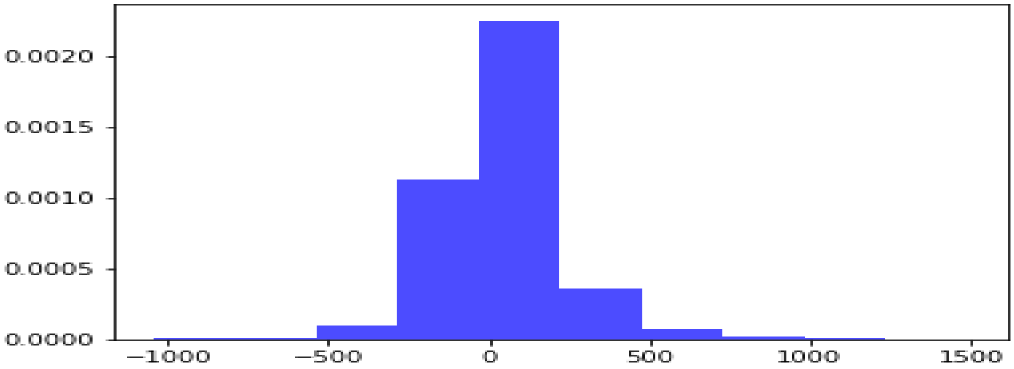
6. If you see count it tells the same record count that is 4100 here. You can see all columns have same count which means there are no missing fields there. You can also check an individual column count, say, for Units , all I have to do is:

7. df['Units'].count()
   Output: 4100

	Week	DayWeek	Day	Month	Hour	Units	Amount
count	4100.000000	4100.000000	4100.000000	4100.000000	4100.000000	4100.000000	4100.000000
mean	34.017805	4.183902	15.812195	8.231463	12.949024	12.779512	35.237046
std	14.714289	1.967864	8.810817	3.396586	2.631853	17.854968	183.538724
min	1.000000	1.000000	1.000000	1.000000	8.000000	1.000000	-1041.800000
25%	29.000000	3.000000	8.000000	7.000000	11.000000	3.000000	-47.000000
50%	37.000000	4.000000	16.000000	9.000000	13.000000	8.000000	-2.385000
75%	45.000000	6.000000	23.000000	11.000000	15.000000	16.000000	99.512500
max	53.000000	7.000000	31.000000	12.000000	19.000000	274.000000	1487.000000

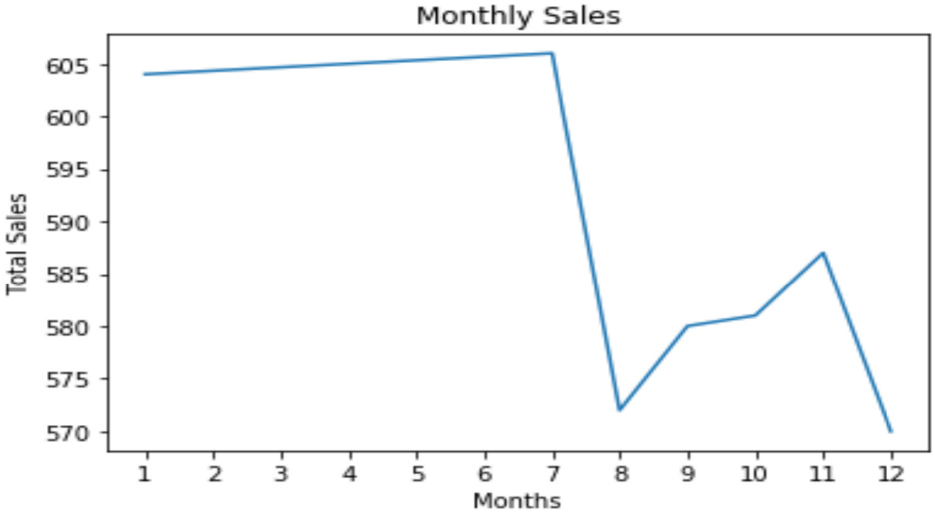
Here we can now picture of how data is available, what is mean, min and max along with standard deviation and median. The percentiles are also there. Standard Deviation is quite useful tool to check how the data is spread above or below the mean. The higher the value, the less is reliable or vice versa. For instance std of Amount is 183.5 while mean is around 35 . On other hand mean of Units is 12.7 and std is 17.85 .

8. Output-Distribution Plot(The data varies from -1000 to 1000 sums up how much the amount varies)Let’s see the distribution of Amount



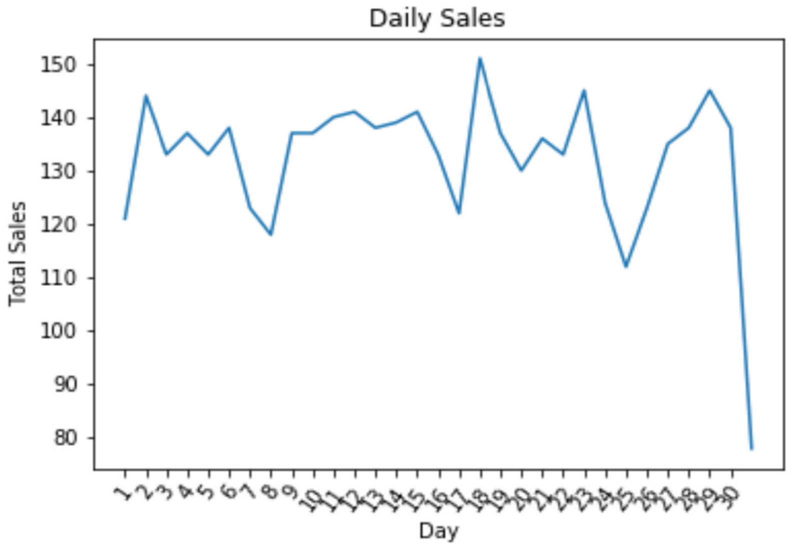
Analysis; Base line which is very large, varies from -1000 to 1000 +

9. Output-Sales by Month



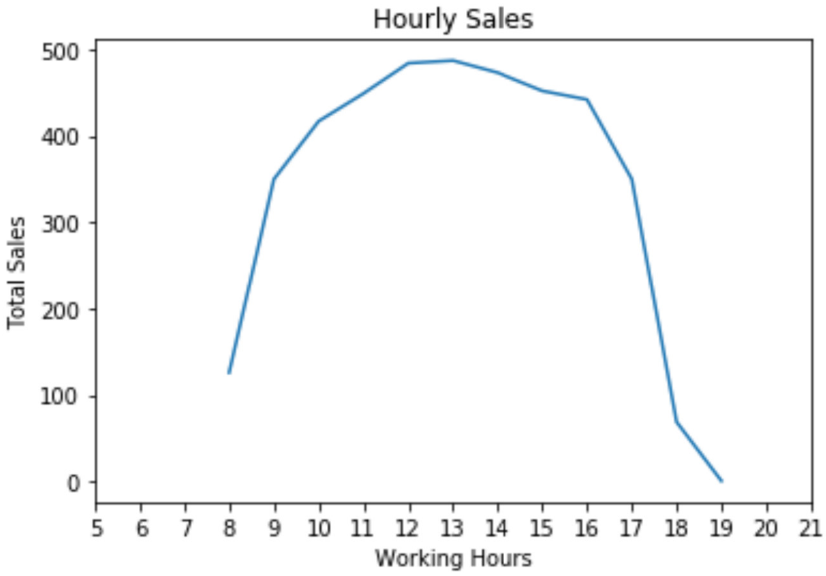
July sales is good but sharp decline in August, then for next 3 months they find it hard to to increase the sales for the last quarter and the years end which is not supposed to happen sales decline again maybe there is no enough sales people to charm customers.

10. Output- Sales By Day



As per this plot, 18th day was the best day as 151 units were sold in that day and sales drastically dropped by the end of the month. May be members get tired or bored? :

11. Output-Hourly Sales



Stores starts around 7 AM in the morning. Majority of the customers come in afternoon. The frequency gets quite low during closing time.