Ampcontrol Assignment
Exploratory Data Analysis for Charging Events

Loc Tran

April 2024

1 Dataset Overview and Descriptive Statistics.

The dataset contains 277 rows. The variables (Column names) are: 'Start Time', 'Meter Start (Wh)', 'Meter End(Wh)', 'Meter Total(Wh)', 'Total Duration (s)', 'Charger_name'. The dataset has the earliest timestamp of 2018-08-24 09:51:00 and latest timestamp of 2019-09-27 12:48:00. Initial overview of the dataset and its variables can be seen below.

	Start Time	Meter Start (Wh)	Meter End(Wh)	Meter Total(Wh)	Total Duration (s)	Charger_name
0	24.08.2018 09:50	50	50.00	0.00	37	NaN
1	24.08.2018 09:51	50	50.00	0.00	38	NaN
2	24.08.2018 09:51	73	118.52	45.52	56	NaN
3	24.08.2018 09:53	105	116.66	11.66	76	NaN
4	24.08.2018 09:54	121	144.77	23.77	19	NaN

(a) The First 5 Entries.

```
RangeIndex: 277 entries, 0 to 276
Data columns (total 6 columns):
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   Start Time          277 non-null    object
 1   Meter Start (Wh)    277 non-null    int64
 2   Meter End(Wh)       277 non-null    float64
 3   Meter Total(Wh)     277 non-null    float64
 4   Total Duration (s)  277 non-null    int64
 5   Charger_name        264 non-null    object
dtypes: float64(2), int64(2), object(2)
memory usage: 13.1+ KB
```

(b) Index Dtype and Columns, Non-null Values.

	Meter Start (Wh)	Meter End(Wh)	Meter Total(Wh)	Total Duration (s)
count	2.770000e+02	2.770000e+02	277.000000	2.770000e+02
mean	3.968875e+05	4.030848e+05	6197.316318	9.651005e+04
std	3.912772e+05	3.892371e+05	12260.182878	3.472706e+05
min	0.000000e+00	0.000000e+00	0.000000	0.000000e+00
25%	6.900900e+04	7.866592e+04	0.000000	1.200000e+01
50%	1.932000e+05	2.007288e+05	1380.280000	5.704000e+03
75%	7.430480e+05	7.508278e+05	6822.500000	7.343900e+04
max	1.204911e+06	1.204935e+06	126350.920000	3.020411e+06

(c)

```
Start Time            0
Meter Start (Wh)     13
Meter End(Wh)         6
Meter Total(Wh)      79
Total Duration (s)   68
Charger_name          0
```

(d) Number of Zeros entries

	Start Time	Charger_name
count	277	264
unique	265	16
top	08.01.2019 13:01	charger_4
freq	4	77

(e) Title for the fifth image

Figure 1: Overall title for all figures

Some observations:

- The variables (columns) with the same unit Wh do not have the same type. **Meter Start (Wh)** is integer while others are floats. We can convert this variable to float for future analysis.
- There are some **NULL** values in **Charger_name** column. We can set this to **'unknown'** for future analysis.
- Start time is in date time format, but it follows **DD/MM/YYYY**. We can change this to **MM/DD/YYYY** for future analysis.
- High standard deviations for **Meter End (Wh)** and **Total Duration (s)** indicate significant variability in the dataset. The Total Duration, in particular, varies greatly from essentially no time to over a month. I will examine this observations further in the Feature Assessment section.
- There are many zeros entries in the dataset, especially **Meter End (Wh)** and **Total Duration (s)**, which might be inactive charger readings or errors. We can consider filtering out these value.
- As seen above, **Meter End (Wh)** has 6 zero entries. Upon further inspection, these entries have **Meter End (Wh)** equals to 0 but has **Total Duration (s)** in thousands which might from error readings. This might contribute to the higher standard deviation of **Total Duration (s)**. We can filter out these entries.
- There are 265 unique timestamp in the datset with some duplicates. We can examine this further to ensure no duplicate readings of the same charger, which indicates errors.

2 Feature Assessment.

All tests and calculations in this section are performed on the raw data with changes to format and N/A handle, which aims to understand the dataset better before perform any further modification. The data report files can be found in the /data_report folder. I also generated data report files for each phase of the experiment (including the final dataset used). The **Meter End (Wh)** is **Meter End (Wh)** - **Meter Start (Wh)** (which is verified in the attached notebooks) so we can focus on **Meter End (Wh)** and **Total Duration (s)**.

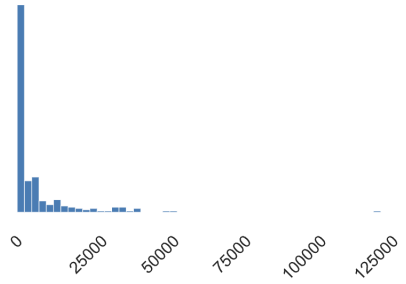
2.1 Meter End (Wh)

Meter Total(Wh)

Real number (ℝ)

ZEROS

Distinct	199	**Minimum**	0
Distinct (%)	71.8%	**Maximum**	126350.92
Missing	0	**Zeros**	79
Missing (%)	0.0%	**Zeros (%)**	28.5%
Infinite	0	**Negative**	0
Infinite (%)	0.0%	**Negative (%)**	0.0%
Mean	6197.3163	**Memory size**	2.3 KiB



More details

Statistics | Histogram | Common values | Extreme values

Quantile statistics

Minimum	0
5-th percentile	0
Q1	0
median	1380.28
Q3	6822.5
95-th percentile	33017.662
Maximum	126350.92
Range	126350.92
Interquartile range (IQR)	6822.5

Descriptive statistics

Standard deviation	12260.183
Coefficient of variation (CV)	1.9783052
Kurtosis	34.845701
Mean	6197.3163
Median Absolute Deviation (MAD)	1380.28
Skewness	4.6608254
Sum	1716656.6
Variance	1.5031208×10^8
Monotonicity	Not monotonic

- **Zeros**: 28.5% of the observations are zeros, which might suggest periods of inactivity or shutdown in the data collection process (errors).
- **Range**: The data has a wide range from 0 to 126350.92 Wh, indicating high variability in the measured energy consumption.
- **Mean vs. Median**: The mean is quite higher than the median (6197.3163 vs. 1380.28), suggesting the data is right-skewed and there are likely outliers that are pushing the mean up.
- **Quartiles**: The 1st quartile (Q1) is at 0, which, along with the high percentage of zeros, indicates that a large number of the values are clustered around 0. The 3rd quartile (Q3) is at 6822.5, which means 75
- **Outliers**: The existence of outliers is also supported by a very high 95th percentile compared to the mean and median, and extreme kurtosis (34.845701) which indicates a high peak and fat tails in the distribution.
- **Standard Deviation**: A large standard deviation (12260.183) relative to the mean suggests that the data points are spread out over a wide range of values.
- **Skewness**: The positive skewness (4.6608254) further confirms that the distribution is skewed to the right, meaning there are a number of very high values.
- **Histogram**: The histogram confirms the skewness and the presence of outliers, with most data concentrated near the lower end and a long tail stretching to the right.
- **Coefficient of Variation (CV)**: The CV is quite high (about 1.98), indicating that the standard deviation is nearly twice the mean, which again underscores the high level of dispersion in the data.

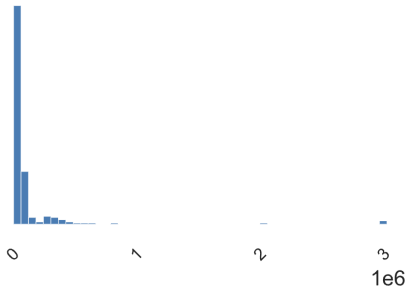
2.2 Total Duration (s)

Total Duration (s)

Real number (ℝ)

ZEROS

Distinct	205	**Minimum**	0
Distinct (%)	74.0%	**Maximum**	3020411
Missing	0	**Zeros**	68
Missing (%)	0.0%	**Zeros (%)**	24.5%
Infinite	0	**Negative**	0
Infinite (%)	0.0%	**Negative (%)**	0.0%
Mean	96510.051	**Memory size**	2.3 KiB



More details

Statistics | Histogram | Common values | Extreme values

Quantile statistics

Minimum	0
5-th percentile	0
Q1	12
median	5704
Q3	73439
95-th percentile	352570.6
Maximum	3020411
Range	3020411
Interquartile range (IQR)	73427

Descriptive statistics

Standard deviation	347270.57
Coefficient of variation (CV)	3.598284
Kurtosis	56.242405
Mean	96510.051
Median Absolute Deviation (MAD)	5704
Skewness	7.2091965
Sum	26733284
Variance	1.2059685×10^{11}
Monotonicity	Not monotonic

- **Zeros**: A significant proportion of the data (24.5%) has a value of zero. This suggests many instances where the duration was not recorded or the event/process did not occur.
- **Range**: The data spans from a minimum of 0 to a maximum of 3,020,411 seconds with a large range, suggesting highly variable durations.
- **Mean vs. Median**: The mean duration is 96,510.051 seconds, which is substantially higher than the median of 5,704 seconds, indicating a right-skewed distribution with some extremely long durations.
- **Quartiles**: The first quartile (Q1) is only 12 seconds, while the third quartile (Q3) is at 73,439 seconds, which indicates that 75% of the durations are less than about 20.4 hours.
- **Outliers**: The very high 95th percentile (352,570.6 seconds) relative to the median suggests that the top 5% of durations are extremely long compared to the rest. Kurtosis value is extremely high (56.242405), which points to the presence of very prominent outliers.
- **Standard Deviation**: The standard deviation is huge (347,270.57 seconds) relative to the mean, indicating that durations vary widely from the average.
- **Skewness**: A very high skewness value (7.2091965) indicates that the distribution is heavily skewed to the right, which is typical for data that includes a lot of very high outliers.
- **Histogram**: The histogram shows that most data points are clustered near the beginning, with a very long tail stretching to the right, further illustrating the right-skewed nature of the distribution.
- **Coefficient of Variation (CV)**: very high CV (approximately 3.59) indicates that the standard deviation is over three times larger than the mean, highlighting the high relative variability in the data.

3 Data Processing

The data processing process can be seen in the notebook file. I made the following changes to the dataset:

- Convert the **Meter Start (Wh)** to float.
- All **NULL** charger names are changed to **unknown**.
- Date now has the form **MM/DD/YYYY**.
- Filter out any invalid entries with neither **Meter End (Wh)** nor **Total Duration (s)** less than 0.
- Add columns for hour in the day and day in the week.

The resulted dataset consists of 175 rows and 8 variables.

4 Visualization.

The weekly and daily patterns can be seen below

4.1 Distribution

Initial distribution of meter total and total duration by hours in a day can be seen below.

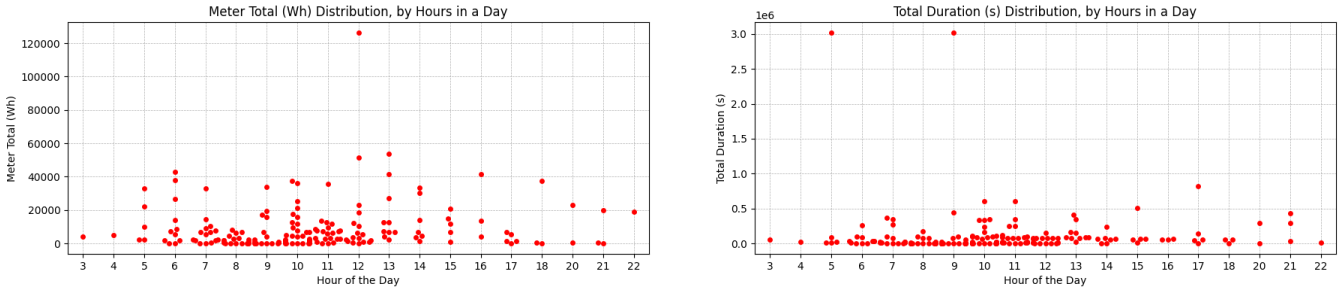


Figure 2: Initial Distribution of Meter Total and Total Duration, by Hours in a Day

We can see that there are some outliers, let's filter them out.

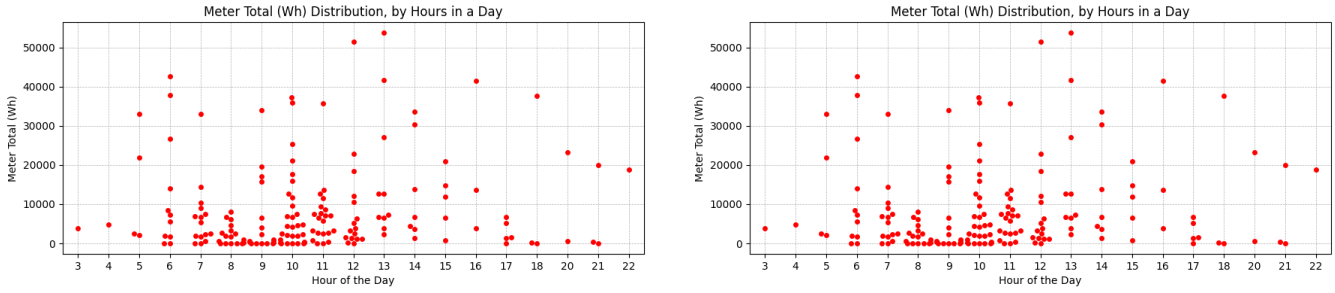


Figure 3: Distribution of Meter Total and Total Duration, by Hours in a Day

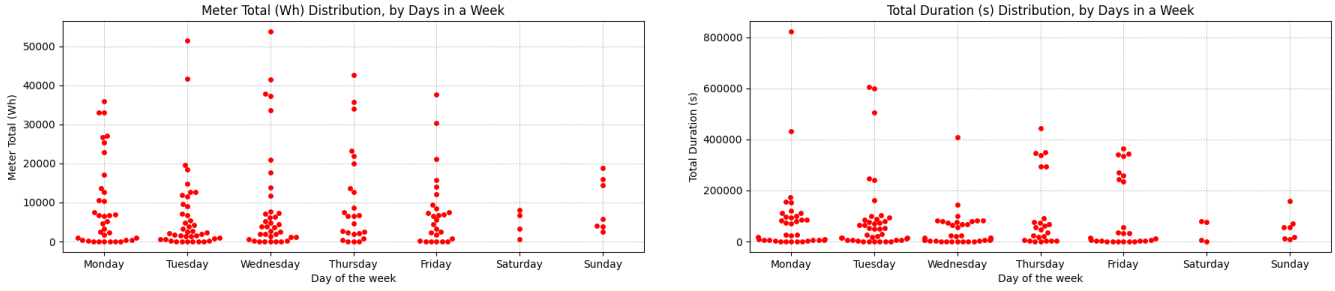
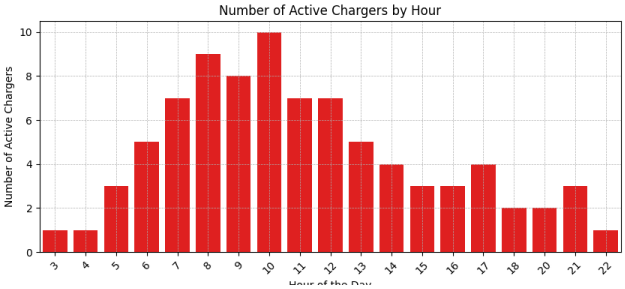


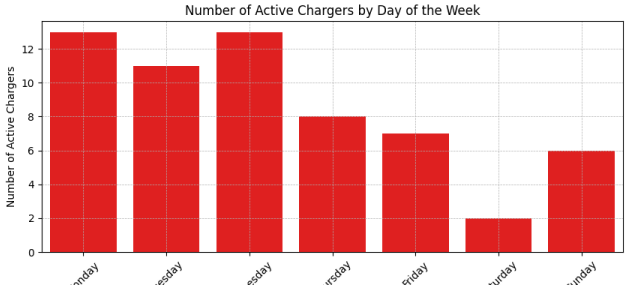
Figure 4: Distribution of Meter Total and Total Duration, by Days in a Week

For both charging duration and total meter, we can see that:

- There is a daily pattern: There are more charging recorded from 7 to 12.
- There is a weekly pattern: There are more charging during the weekday.

We have the following histograms of active charger to further visualize the pattern.



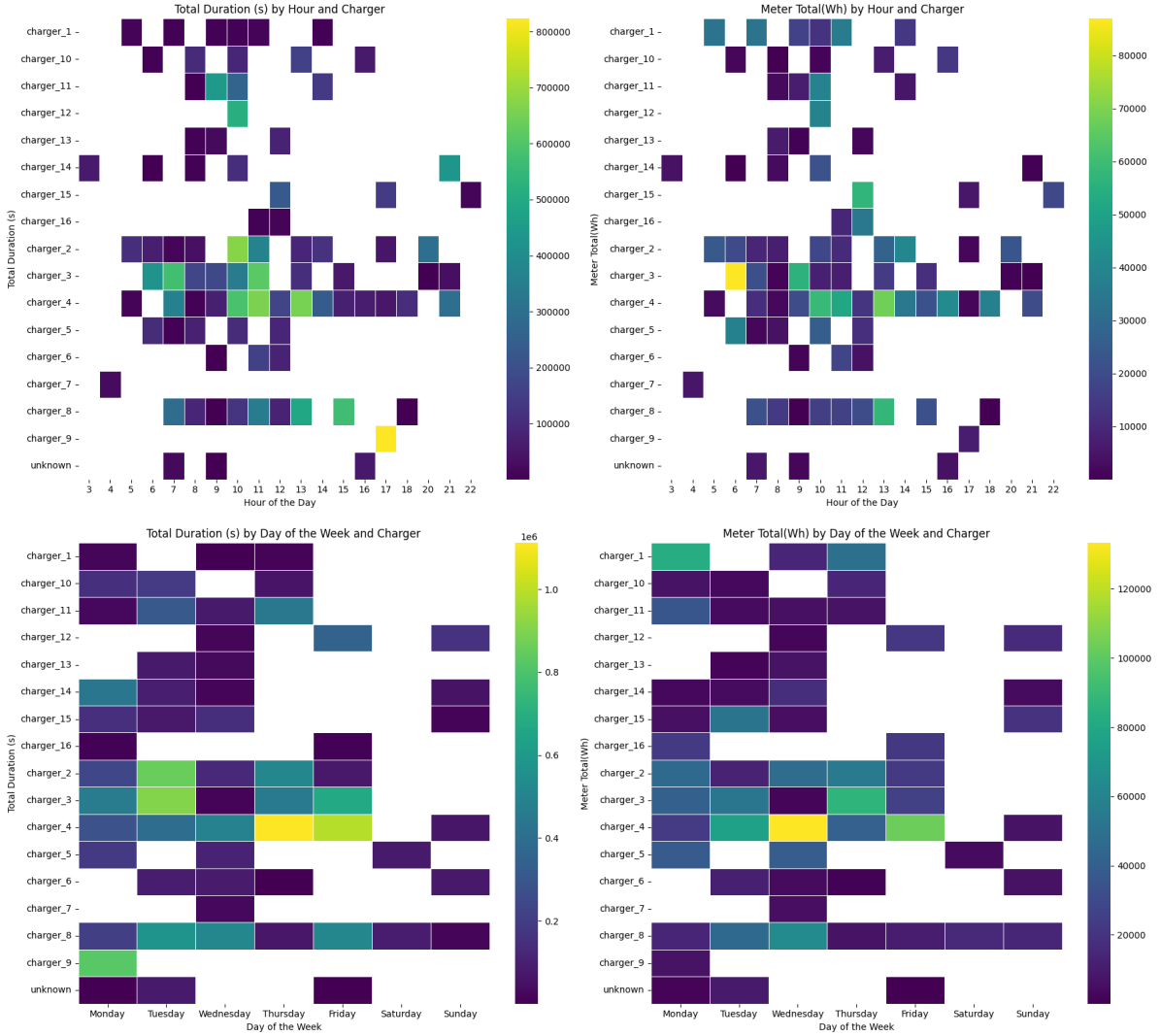


Observations:

- From the hourly distribution chart, there's a noticeable peak in the number of active chargers during the late morning and early afternoon hours.
- The daily distribution chart shows a higher number of active chargers towards the beginning of the week, with the peak on Tuesday, followed by a gradual decrease through to Sunday. This pattern could suggest that charging activity is higher during the working days.

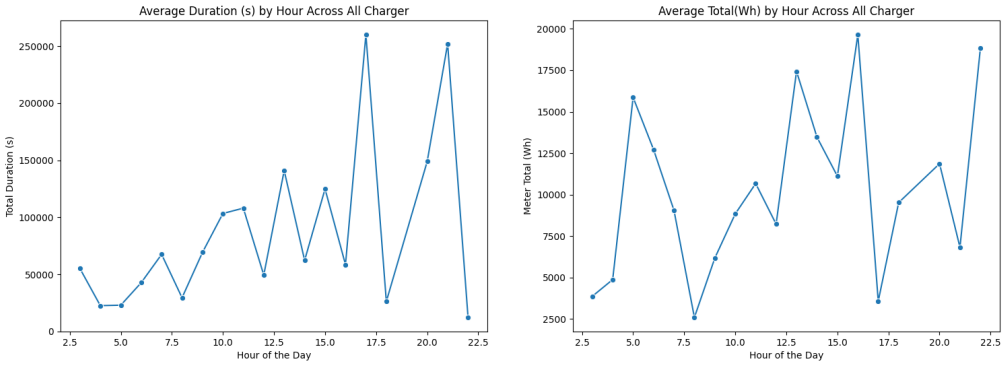
.

4.2 Heat Maps



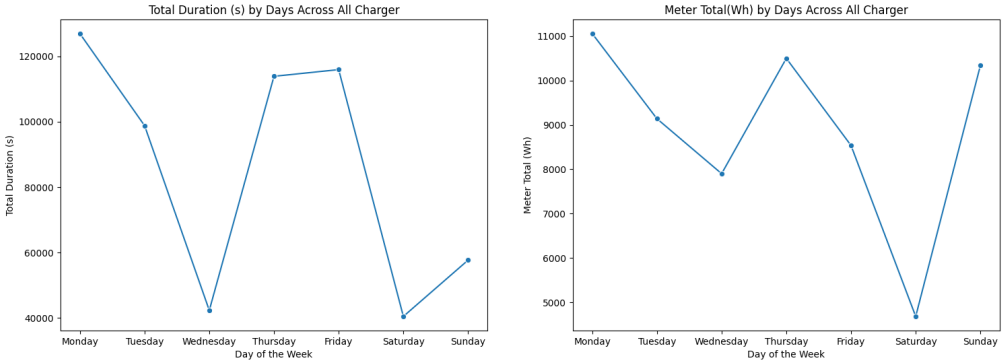
The daily and weekly patterns are shown clearly in the heat maps. We can also see the workload imbalance between all chargers (both across hours and days).

4.3 Average Line Graph



Some observation:

- This graph indicates variability in charging duration throughout the day, with significant spikes at certain hours, particularly towards the evening and night. There's a notable peak at around 13:00 and a higher peak after 20:00, suggesting that users tend to charge for longer durations during these hours. It is noted that even with fewer active chargers at these times, the charging sessions are longer, which could indicate overnight charging.
- The power usage, measured in watt-hours, also fluctuates throughout the day. The trend doesn't appear to be as clear-cut as the duration, but there are noticeable increases at several points in the day, with peaks occurring during the night and early morning hours.



- The graph indicates substantial variation in total charging duration by day of the week. There are peaks on Tuesday and Friday, with a significant drop on Wednesday, suggesting inconsistent charging behavior across the week.
- Similar to the total duration, the total energy consumption (meter total in watt-hours) shows fluctuations throughout the week. There's a sharp decrease on Wednesday, followed by a significant increase on Thursday, and again a drop on Sunday.

5 Data Quality Evaluation.

Even though daily and weekly pattern can be seen with both Meter Total and Total Duration, we have to consider the following drawbacks of the dataset:

- There is no information regarding the variables in the dataset (How some variables are recorded?). Insights on this might explain variables like Total Duration, Meter Total better due to its variance (Some are in thousands while some are in lower digits).
- Large number of error readings.
- Large number of readings for certain chargers while small number of readings for others. Some chargers have duplicate readings or less than 2 readings.
- Infrequent data (both in time series and charger wise). Additional data from others charging network (in the same considered geography location) can reduce bias.

With all these drawbacks and the limited amount of data, biases could happen during evaluation. Further analysis of Charging Duration and Meter in Hourly/Weekly, Weekdays/Weekend, Summer/Winter, Holiday Month/Non Holiday Month can be done to find other patterns in the future with the addition of variables (created from current dataset or external dataset) and sufficient number of data.