BI Final- 19667

Tool Used: PowerBI

Dataset:

https://www.kaggle.com/datasets/taranvee/smart-home-dataset-with-weather-information

The following points were summarized about the data after the data understanding phase:

1. The Dataset is of a smart home service product which is tasked with recording measurements/readings.
2. The data describes the electricity consumption of house components, which are appliance in use and power usage in different rooms, in kilo-Watts.
3. The dataset also has information related to the electricity generation being done by use of solar power, also in [kW].
4. After describing the consumption by the aforementioned components, the dataset gives extensive weather-related information such as:
   i) Categorical: How is the weather? Sunney? Rainy? Etc.
   ii) Numerical: Weather indicators such as Temperature, humidity index, etc.
   iii) Precipitation index and expectation
5. The columns related to weather information usually converge to give us the weather condition (Summary) and temperature, so we'll use these two columns for our primary analysis.

Problem Statement(s):

1. Make the Electricity Consumption and generation efficient using smart home technology.

2. Control the electricity consumption and bringing it down by focusing on individual elements like appliances/rooms, weather conditions, usage over time, etc.

Data Wrangling:

Preparation and Transformation:

- After understanding the data, it was observed that the dataset had 32 columns and almost 500K rows.
- The Dataset was loaded to python, and we observed the rows and columns.
- Based on the wrangling technique taught in the course, the wrangling started in the following manner
- Observed the shape of data, head and tail of data rows and data types of the fields present.
- Missing Value replacement:
  1. While observing the tail and shape, we saw that there was a '/' in one row and no corresponding data
  2. We used isna().sum() function to calculate missing value, and it generated 1 in all fields except one, thus confirming our assumption of having the slash in one row.
  3. Best strategy to handle the missing value here was the deletion on that 1 row because deleting it will have no effect on the data, so we did it.

Missing Values Analysis

```
In [8]: #Dealing with null values
        print(df.isna().sum())
        msno.bar(df)
        #The slash in the last row is causing the only mssing value in the dataset, so we remove it
```

```
In [9]: df=df.dropna()
```

```
In [10]: print(df.isna().sum())
         msno.bar(df)
         #Missing values replaced
```

```
time                  0
use [kW]              0
gen [kW]              0
House overall [kW]    0
```

- Taking care of the time column, which is the most important column for analysis:
  - Time column had values as big integers, and the dataset description told us that the time was recorded by a 1-minute interval of first 350 days of 2016.
  - The pandas function 'to_DateTime()" does work on an integer range sso big, so it couldn't be used.

```
In [12]: #df['time']=pd.to_datetime(df['time'], unit='m')
         #your_timestamp = 1331856000000
         df['time']=datetime.datetime.fromtimestamp(df['time'] / 1e3)
         #df['time'] =  pd.to_datetime(df['time'], format='%d%b%Y:%H:%M:%S.%f')
```

```
---------------------------------------------------------------------------
TypeError                                 Traceback (most recent call last)
~\AppData\Local\Temp/ipykernel_9960/2383959152.py in <module>
      1 #df['time']=pd.to_datetime(df['time'], unit='m')
      2 #your_timestamp = 1331856000000
----> 3 df['time']=datetime.datetime.fromtimestamp(df['time'] / 1e3)
      4 #df['time'] =  pd.to_datetime(df['time'], format='%d%b%Y:%H:%M:%S.%f')

C:\ProgramData\Anaconda3\lib\site-packages\pandas\core\series.py in wrapper(self)
    183         if len(self) == 1:
    184             return converter(self.iloc[0])
--> 185         raise TypeError(f"cannot convert the series to {converter}")
    186
    187     wrapper.__name__ = f"__{converter.__name__}__"

TypeError: cannot convert the series to <class 'int'>
```

  - Instead of using the function mentioned above, we replace the column by a column of **our own,**
  - In this column that we make for Time, we will make our own time series accurately replicating the original one and will adjust it according to the description given in the description of dataset,
  - We start, according to description, the time from 1st Jan 2016 at 5:00 AM, and apply a 1-minute increment on every row, thus achieving an accurate result.
  - We discard the original time column and replace it with the new one.

```
In [13]: #Since none of the methods worked using the to_datetime function
         #The technique was to acquire knowledge of time of readings from the online
         # link, then apply the sequence to data set. Kaggle link said the the readings were
         #from 1 jan 2016  to 15th dec 2016 (350 days) so thats what we try to achieve
         #the conversion through a different technique by using a unix epoch method to make the column a
         #time series of our own which matches the given range
         time=pd.date_range('2016-01-01 05:00:00', periods=503910, freq='min')
         time=pd.DatetimeIndex(time)
         df['time']=time

In [14]: df['time']

Out[14]: 0         2016-01-01 05:00:00
         1         2016-01-01 05:01:00
         2         2016-01-01 05:02:00
         3         2016-01-01 05:03:00
         4         2016-01-01 05:04:00
                            ...
         503905    2016-12-16 03:25:00
         503906    2016-12-16 03:26:00
         503907    2016-12-16 03:27:00
         503908    2016-12-16 03:28:00
         503909    2016-12-16 03:29:00
         Name: time, Length: 503910, dtype: datetime64[ns]
```

- Next step is to take care of the “cloud cover” column.
- Although we didn’t use this column in our analysis, we should take care of this column before proceeding to data analysis.
- We delete the string variables in this column to null values, then used backfill to deal with these values to replace it all the way to the top.

```
In [15]: df['cloudCover']= df['cloudCover'].replace('cloudCover', np.nan)

In [16]: print((df['cloudCover'].isnull().sum()/df.shape[0])*100, '%\n')
         #We have printed to see how much of the whole column had 'CloudCover' as a value
         #whereas cloudCover is a numeric column
         #We use backfill to replace the string value from this column
         df['cloudCover']=df['cloudCover'].fillna(method='bfill')
         df['cloudCover']=df['cloudCover'].astype(float)
         #Backfill takes the valid value and takes it to replace all the way to the first row.

         0.01150999186362644 %

In [17]: print(df['cloudCover'].value_counts())

         0.00    68236
         0.31    49899
         1.00    48705
         0.03    33940
```

- ‘Icon’ and ‘Summary’, ‘Solar [kW]’ and ‘gen [kW]’ and ‘use [kW]’ and ‘House Overall [kW]’ are the pairs of columns that are the same, so we drop one of these columns from the mentioned pairs.

```
In [18]: df['House overall [kW]'].equals(df['use [kW]'])
         #Answer=true means that both columns are exactly the same

Out[18]: True

In [19]: df['Solar [kW]'].equals(df['gen [kW]'])
         #Answer=true means that both columns are exactly the same

Out[19]: True

In [20]: df.drop('use [kW]', axis =1, inplace=True)
         df.drop('Solar [kW]', axis =1, inplace=True)

In [21]: df['icon'].equals(df['summary'])
         #Answer=false means that both columns are not the same
         #We still want to drop the icpn column because:
         #1. it is changing exactly as summary column
         #2. Summary column does the same thing which icon column does, do its redundant to have both

Out[21]: False

In [22]: df.drop('icon', axis =1, inplace=True)
         df.shape

Out[22]: (503910, 29)
```

- There were no cost columns in the dataset, whereas cost is one of the biggest indicators of consumption in such datasets,

- We create the costs columns by multiplying the total consumption of house by the **KE electricity price kW/minute (0.277 PKR) in 2016**.

In [23]:
```
#Worth noticing that we have no cost or price expenditure indicators in the dataset.
#So we introduce a column which tells us the price of electricity being spent instantaneously.
#We get this column by taking the electricity price and multiplying it to the total electricity used.
#for analysis, we use KE data. According to KE online website https://nepra.org.pk/tariff/Tariff/KESC/2019/SRO%20576%20I%202019%2
#, the price for 1 kW per minute in 2016 was 0.27166 ruppees
df['Usage Electricity']=0.27166*df['House overall [kW]']
```

In [24]:
```
#since we are also generating electricity using solar power, it
#would make sense to see how much electricity we are making too
df['Electricity generated']=0.27166*df['gen [kW]']
```

In [25]:
```
df.head()
```

Out[25]:

Wine cellar [kW]	Garage door [kW]	...	apparentTemperature	pressure	windSpeed	cloudCover	windBearing	precipIntensity	dewPoint	precipProbability	Usage Electricity	Electricity generated
0.006983	0.013083	...	29.26	1016.91	9.18	0.75	282.0	0.0	24.4	0.0	0.253414	0.000946
0.006983	0.013117	...	29.26	1016.91	9.18	0.75	282.0	0.0	24.4	0.0	0.253821	0.000942

- We use box plots to analyze outliers. There were no apparent outliers in the data since consumption varies highly in extreme seasons, however there were a few instances when the consumption was very high and were back to normal the next minute, so it was decided to remove these rows as they were very rare instances and they could distort the analysis.

In [26]:
```
df = df[df['House overall [kW]'] < 12] # Removing rows with high electricty spending because it was an uncommon occurence
plt.figure(figsize=(12, 3), dpi = 100)
sns.boxplot(df['House overall [kW]'])

df.shape
#  37 ROWS removed
```

```
C:\ProgramData\Anaconda3\lib\site-packages\seaborn\_decorators.py:36: FutureWarning: Pass the following variable as a keyword
rg: x. From version 0.12, the only valid positional argument will be `data`, and passing other arguments without an explicit
yword will result in an error or misinterpretation.
  warnings.warn(
```

Out[26]: (503844, 31)

- Inferential Analysis:
  Since we have maximum numerical data and only a small number of categorical fields, and it is clear that appliances will not have a positive or negative correlation with each other since they are used independently, we only run anova test to confirm some of our assumptions about the data that only weather will impact the total generation or consumption of electricity. All the other correlations are just a circumstance of this. See below:

ANOVA

Bewteen one numerical and other categorical variable Here the best columns to run inferential analysis on are:

1. "House Overall" and "Summary' to confirm our assumption which is that changing weather conditions will affect the electricity consumption
2. Electricity generation "gen [kW] "and weather conditions "summary" to see how the electricity generation changes over changing weathers

In [29]:
```
import statsmodels.api as sm
from statsmodels.formula.api import ols

y=df['House overall [kW]']
model = ols('y ~ C(Q("summary"))', data=df).fit()
anova_table=sm.stats.anova_lm(model, typ=2)
display(anova_table)
```

	sum_sq	df	F	PR(>F)
C(Q("summary"))	1412.866490	17.0	75.60828	1.434776e-262
Residual	553813.348465	503826.0	NaN	NaN

In [30]:
```
y=df['gen [kW]']
model = ols('y ~ C(Q("summary"))', data=df).fit()
anova_table=sm.stats.anova_lm(model, typ=2)
display(anova_table)
```

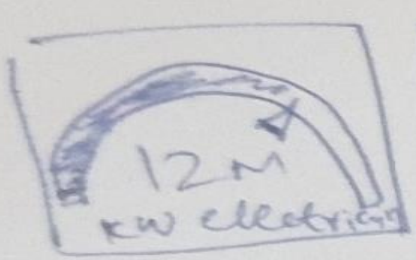
	sum_sq	df	F	PR(>F)
C(Q("summary"))	32.163842	17.0	115.143767	0.0
Residual	8278.642629	503826.0	NaN	NaN

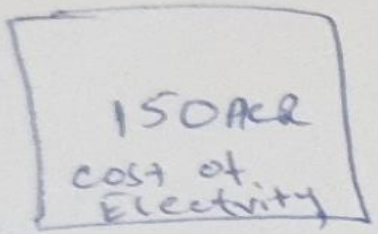
Strategy behind Data Transformation:

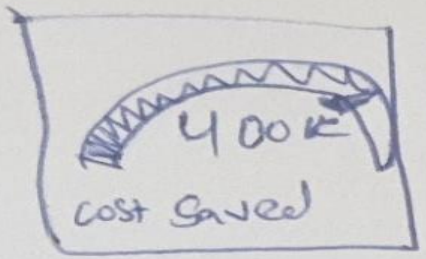
Note:
Electricity usage by appliances and rooms is in kilowatts.
Cost of electricity is in PKR according to KE electricity prices (2016).
'Summary' is a consequence of all the weather-related columns.

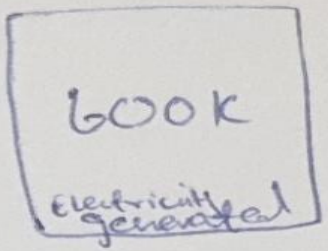
The strategy used here is that we have transformed the data and made necessary changes to it to make it fool-proof when we begin analyzing the electricity usage and how it can be made efficient. We include major factors that are not present i.e. cost of electricity and dropped unwanted columns. The transformed data itself tells us most about how the analysis will be.

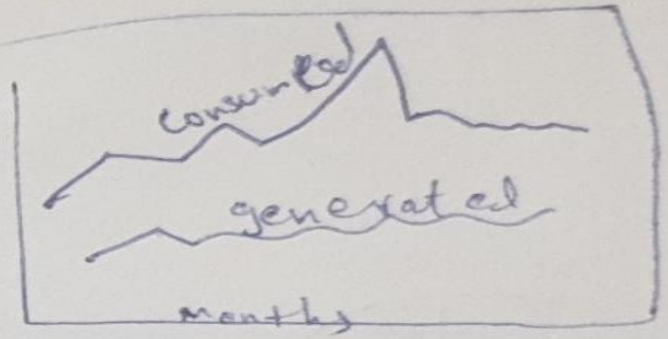
Dashboard 1

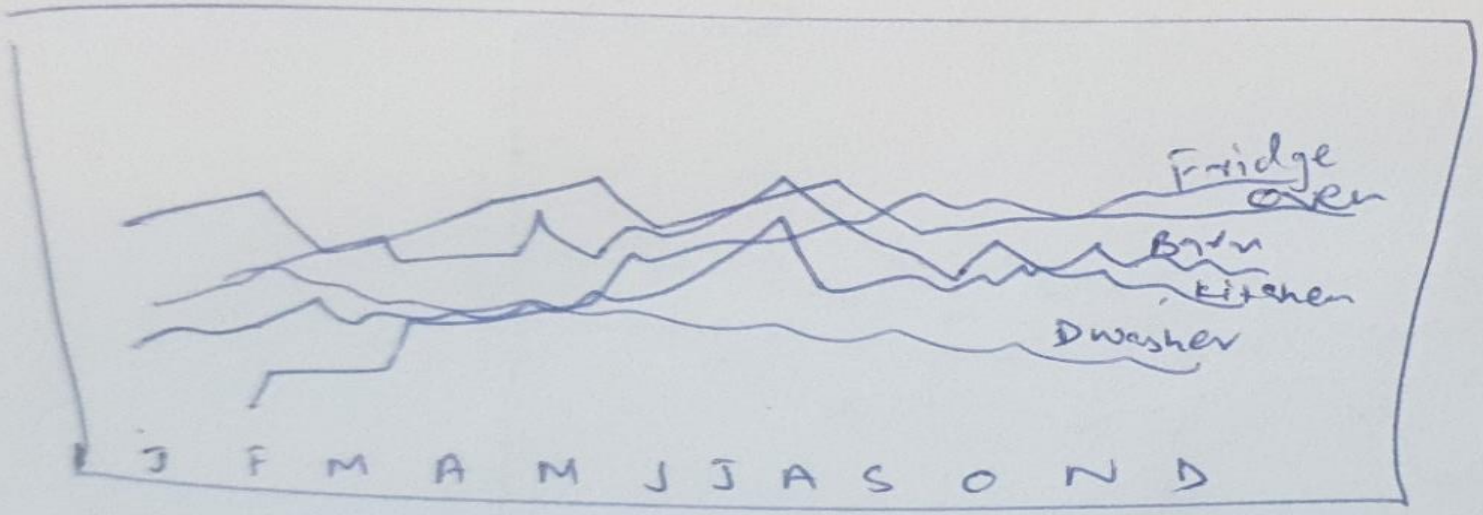

12M
KW electricity

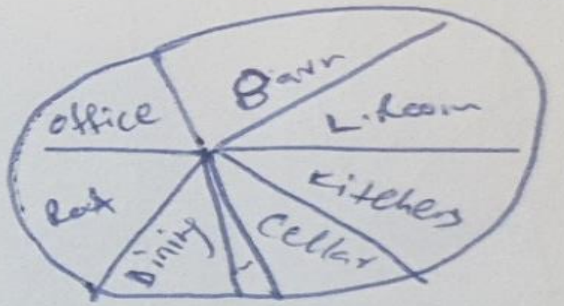

150Ksh
cost of
Electricity

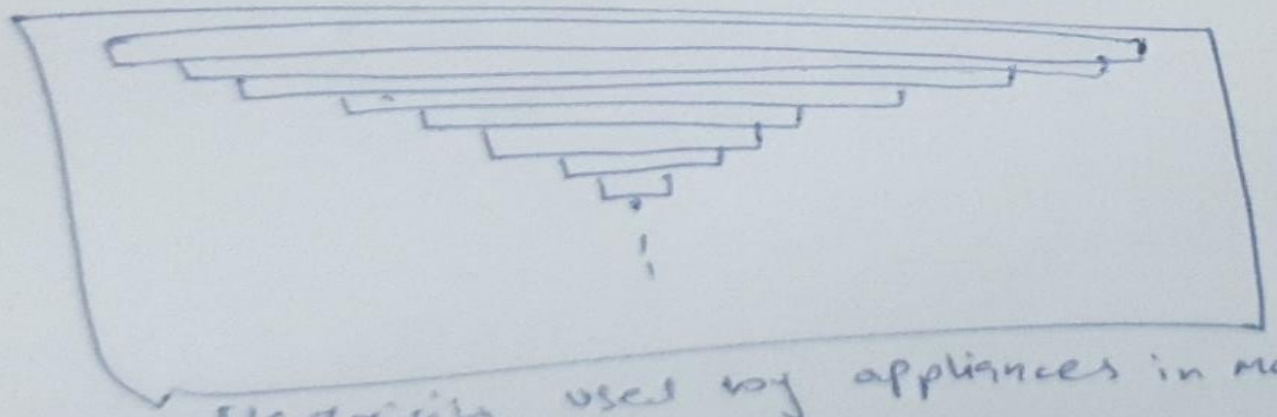

400K
cost saved

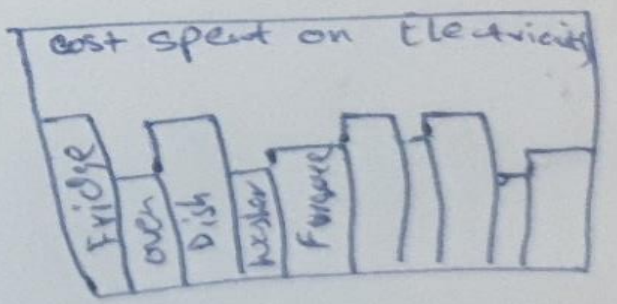

600K
Electricity
generated


consumed
generated
monthly

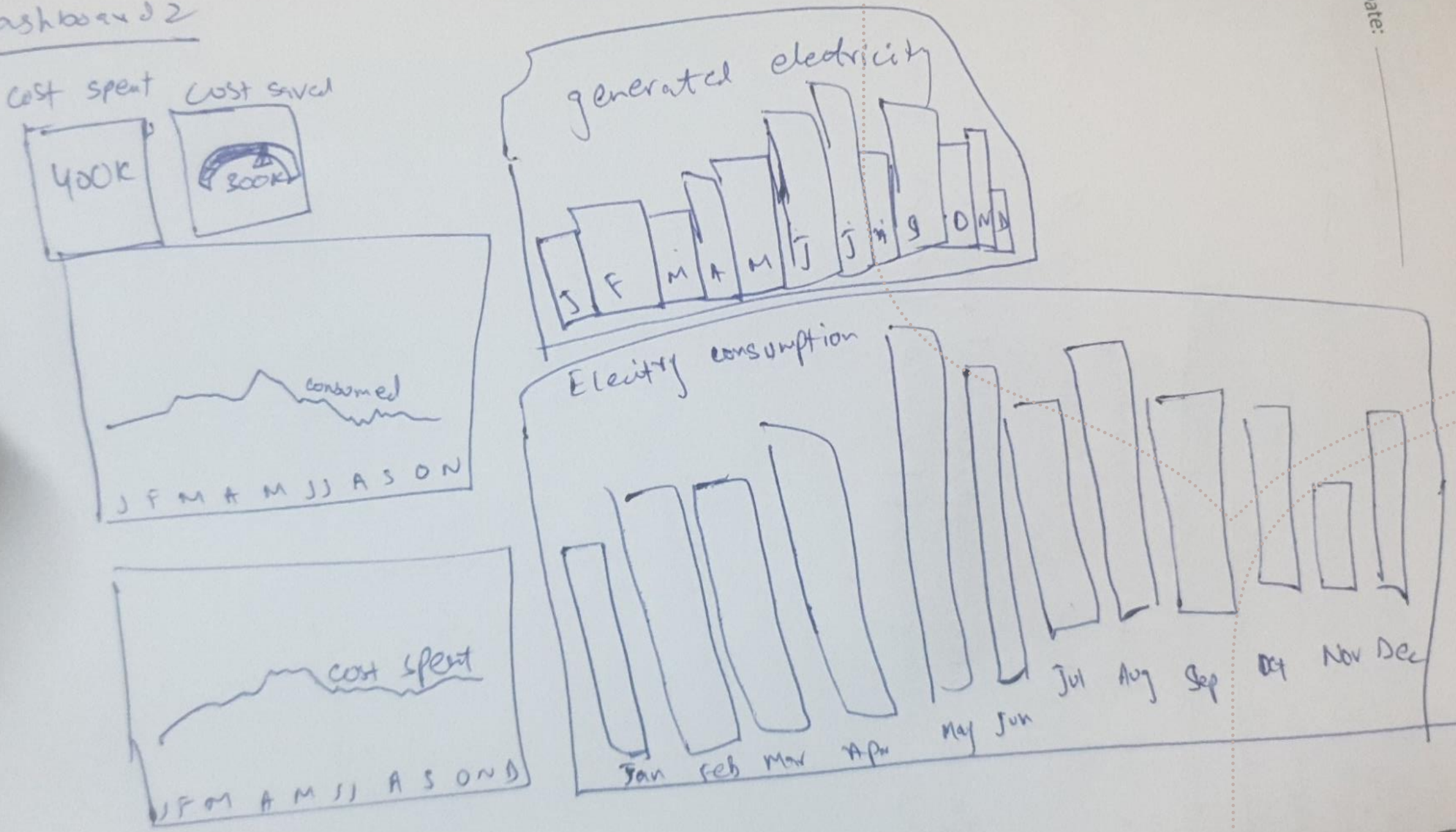

Fridge
oven
Bath
Kitchen
Dwasher
J F M A M J J A S O N D


office
Bedr
L.Room
Bath
Kitchen
Dining
Cellar


Electricity used by appliances in months


cost spent on Electricity
Fridge
oven
Dish
Washer
Fireplace

Dashboard 2



Date:

Dashboard - 3

Days 1, 2, 3 ...

Choose Month

- Jan
- Feb
- March
- April
- May
- June
- July
- Aug
- sept
- Oct
- Nov
- Dec

- Q1
- Q2
- Q3
- Q4

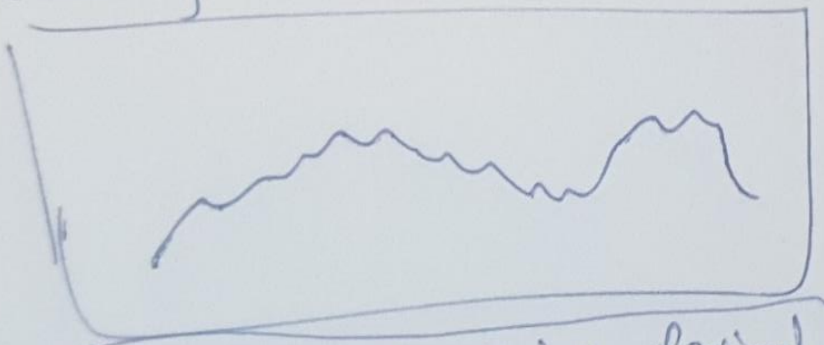
weather during period

Sunny

Electricity usage

	Living	Dining
Bar	Office	Lounge
Dem		

Electricity consumed in Period



Temperature

36°C

Consumption in kW

40 kW

cost spent during Period

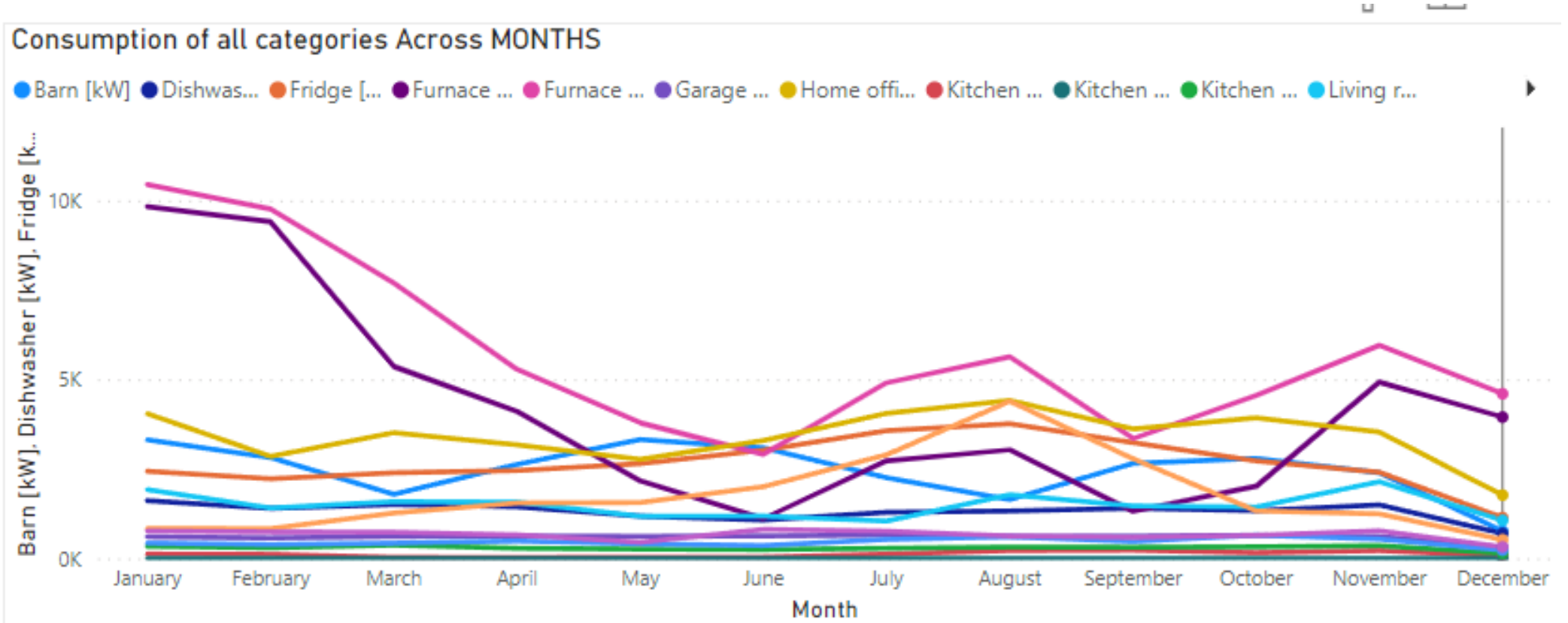
wind speed

42 kph

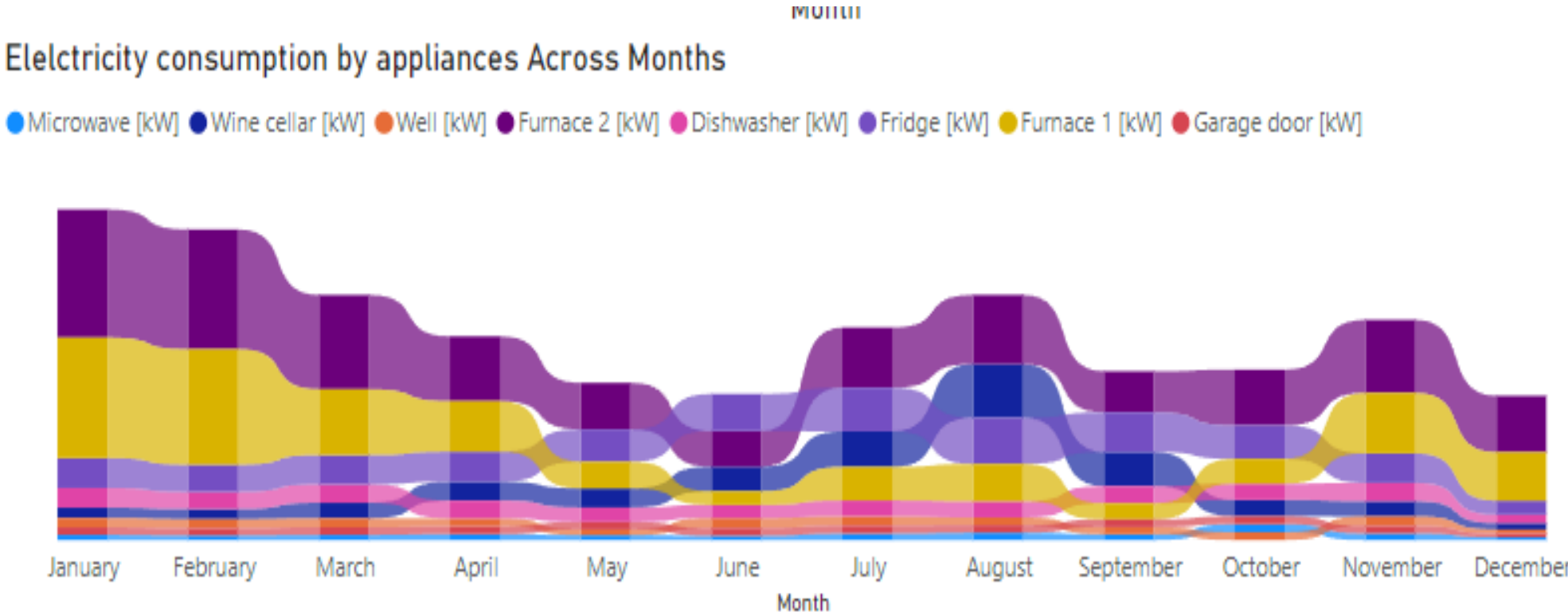
generation in kW

20 kW

Date: _____

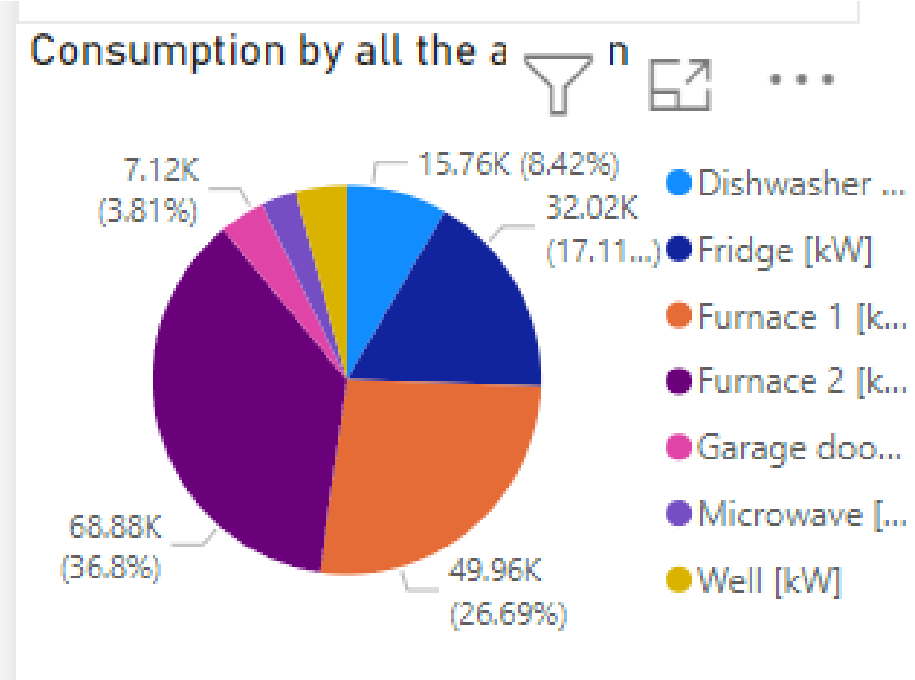


The chart above is a line chart describing one of the most important basic analysis. This charts describes the usage or consumption of electricity by ALL the categories (Rooms and appliances) in each month. Since the dataset only covers 2016, it was best to analyze in Months.
The chart shows that Furnaces utilize the most electricity whereas home office and living room are the leading runner ups.

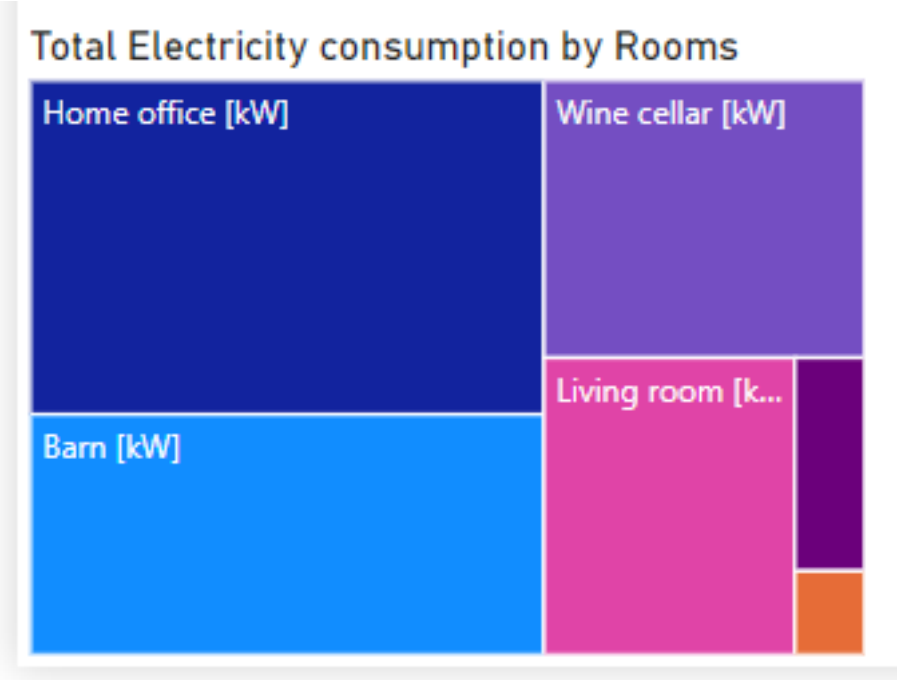


The chart above is a Ribbon chart describing another important basic analysis. This charts describes the usage or consumption of electricity by ALL the APPLIANCES during each month. Since the dataset only covers 2016, it was best to analyze in Months.

The chart shows that Furnaces utilize the most electricity whereas fridge and dishwasher are the leading runner ups. It also describes how the usage increases or decreases during different seasons/weathers.

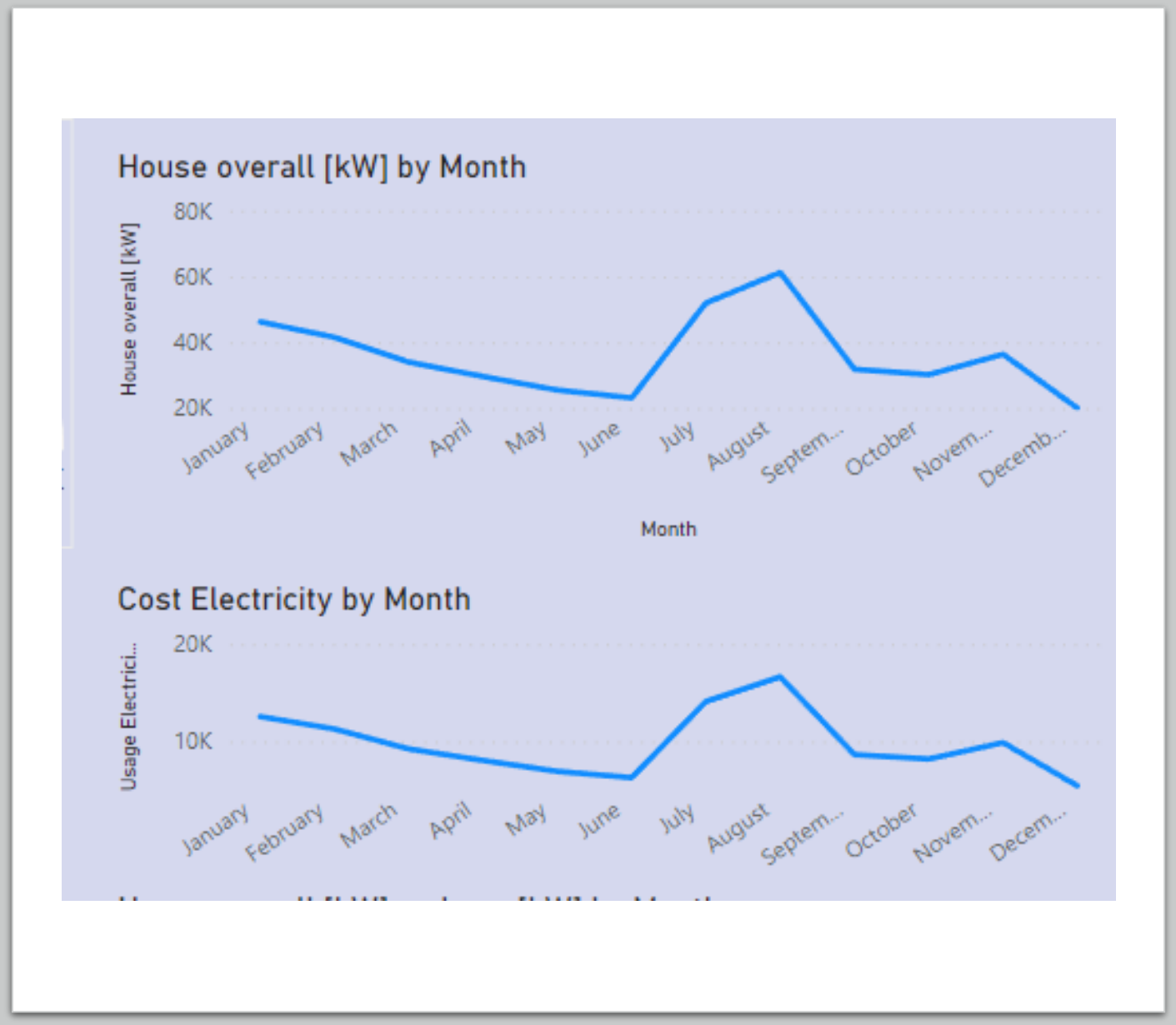


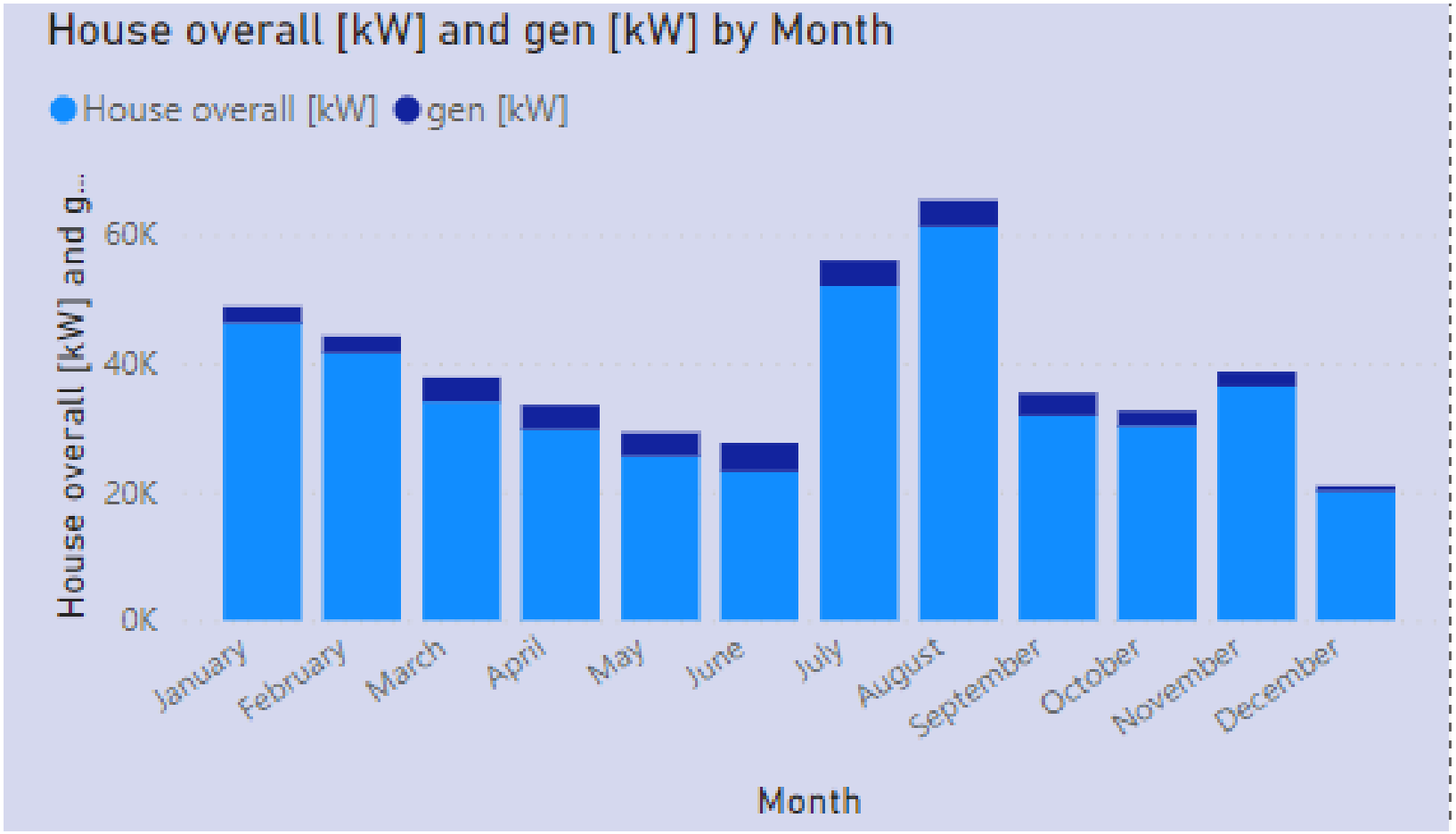
The Pie chart above describes the breakup
Of how much every appliance contributes
Towards the total electric consumption. Furnaces
And fridge are the top contributors.



The Treemap above describes the breakup
Of how much every appliance contributes
Towards the total electric consumption.
Home office, Barn and living room are the
top 3 contributors.

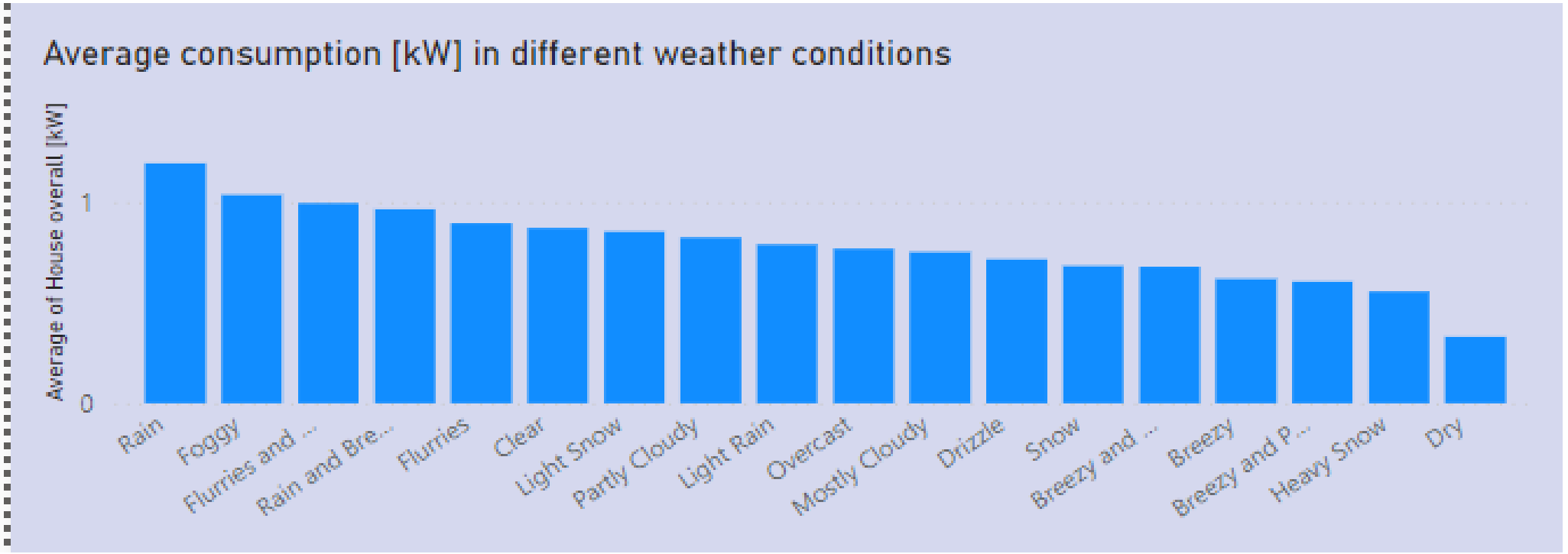
- The charts here are line charts and even though they look similar, they are important towards our analysis because they correlate two major things. The charts show the electricity consumption during each month and consequently the cost spent on electricity during 12 months of 2016. We see that both these measures increase as summers approach and then go down in the winters.





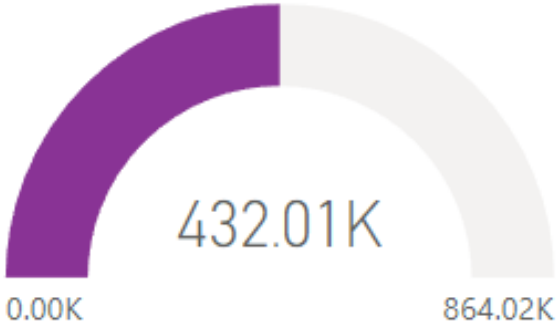
The chart above is a Stacked Bar chart comparing the kW electricity consumed and generated during different months of 2016.

We see by this chart that the generation of electricity through solar power or similar means is extremely low compared to the electricity the house consumes. Generation is also low because we saw that the weather conditions in the area where the data is recorded is mostly rainy and cloudy, thus the solar generation is underproduced.

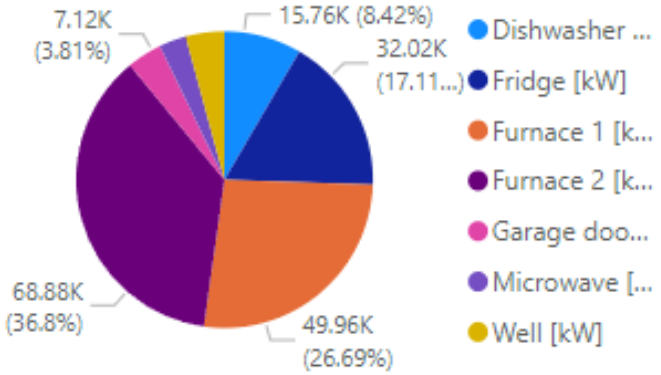


The chart above is a Stacked Bar chart describing the average electricity consumed during different weather situations throughout the year.
As noticed in real life, the electricity consumption is very high during extreme temperatures and low when the temperatures are in the middle, during seasons like autumn and springs when the weather's dry or breezy.

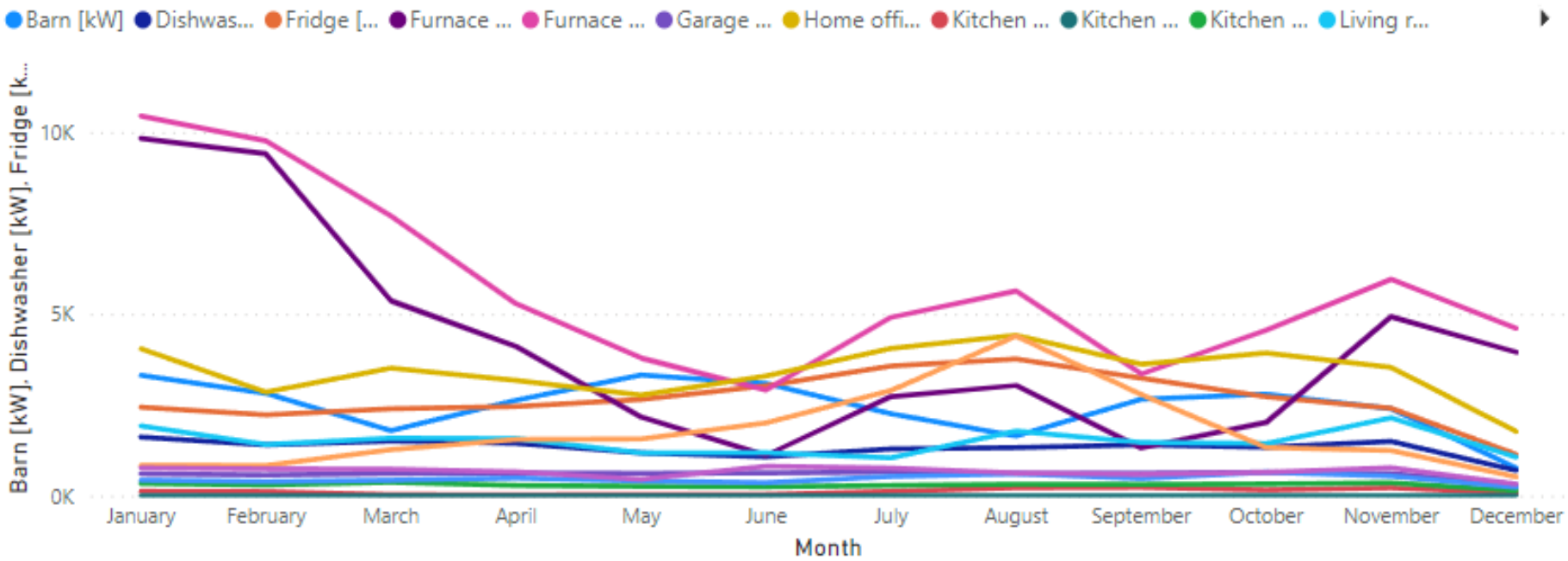
Total Electricity Consumption in [kW]



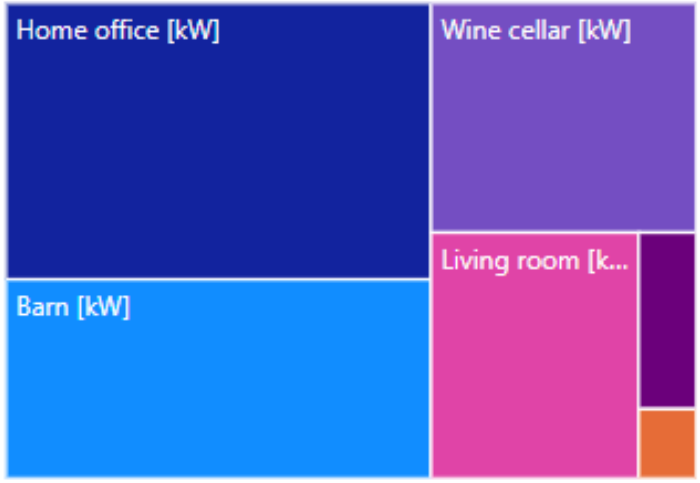
Consumption by all the appliances



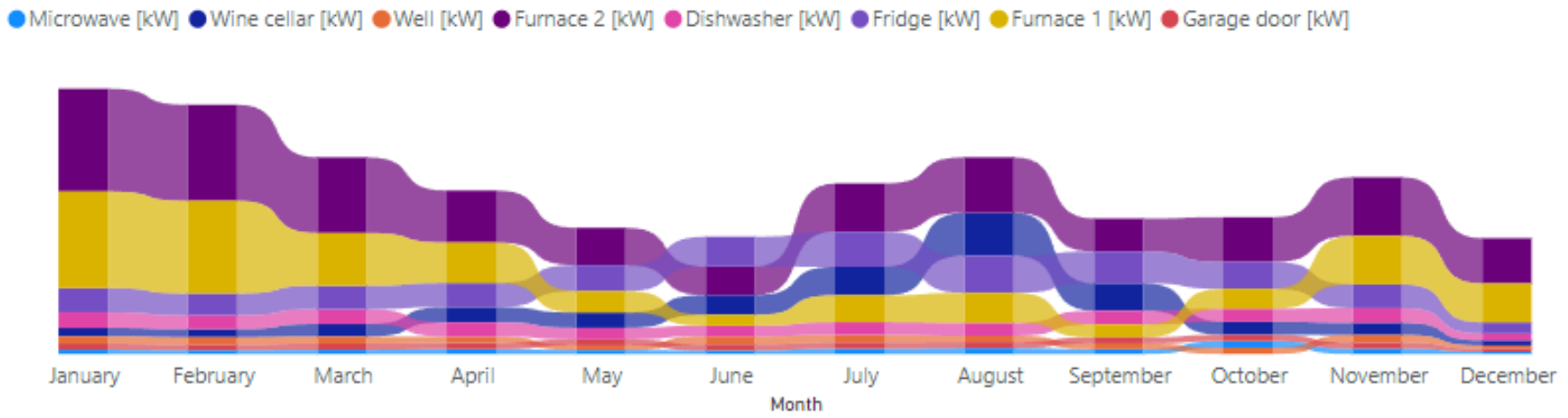
Consumption of all categories Across MONTHS



Total Electricity consumption by Rooms



Elelctricity consumption by appliances Across Months



Analysis of Dashboard: Appliance Consumption(1/2)

- The dashboard summarizes the following analysis from the given data:
- Visualizes the electricity consumed by all the categories throughout different months of 2016 (Line Chart).
- Demonstrates how much electricity is consumed by every appliance throughout 2016, this visualization also compares the consumption of appliances and shows the trends over different seasons/times. (Ribbon Chart).
- Shows how much of the total consumption is taken by each appliance in the form of a Pie chart.

Analysis of Dashboard: Appliance Consumption(2/2)

- Shows how much of the total consumption is taken by each appliance in the form of a Pie chart.
- Shows how much electricity is consumed in each room and how much each room takes up from the total electricity consumption. (Treemap)
- Shows the total kilo-Watts electricity consumed in 2016.

Month

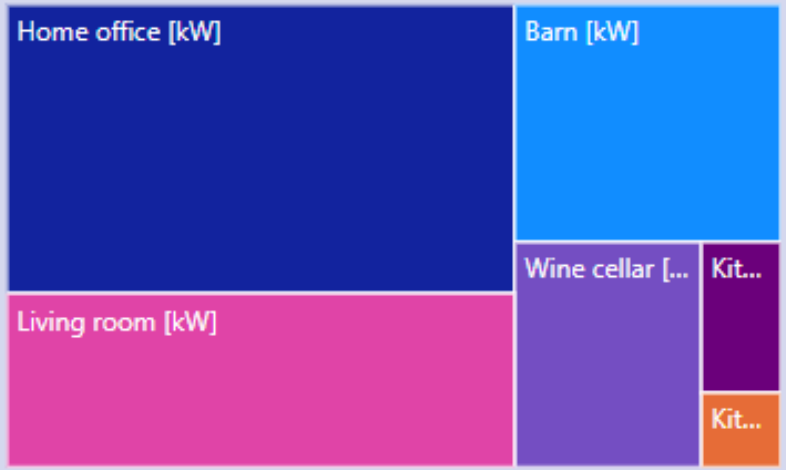
- July
- August
- September
- October
- November
- December

Quarter

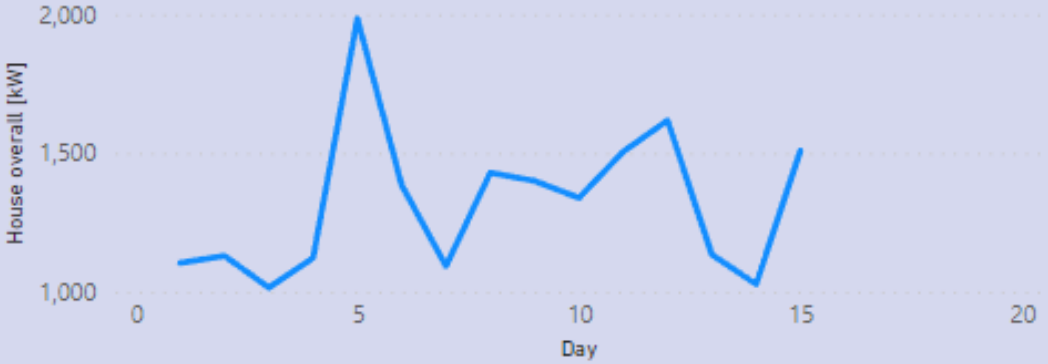
- Qtr 4

0 15

Total Electricity consumption by Rooms



Electricity Consumed [kW] by Day

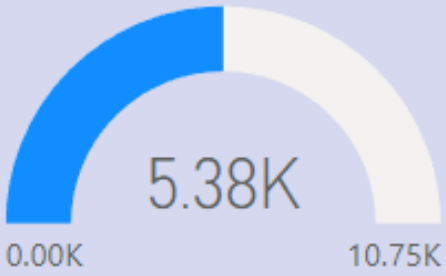


kW Electricity Consumed

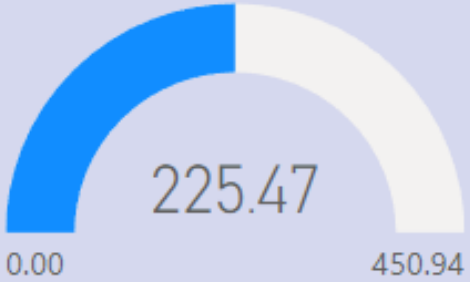
19.79K

House overall [kW]

Cost spent on Electricty in PKR

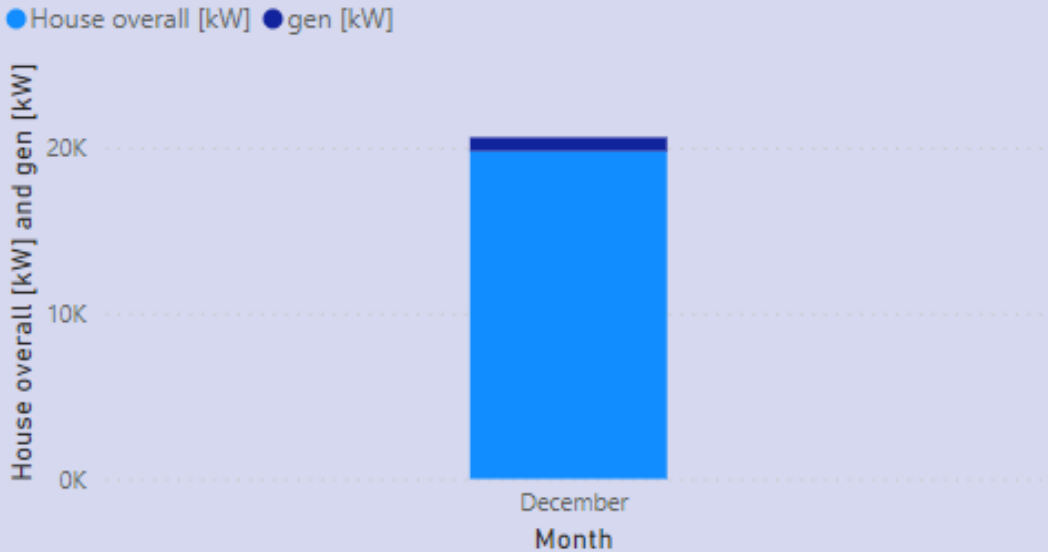


Cost saved on Electricity in PKR

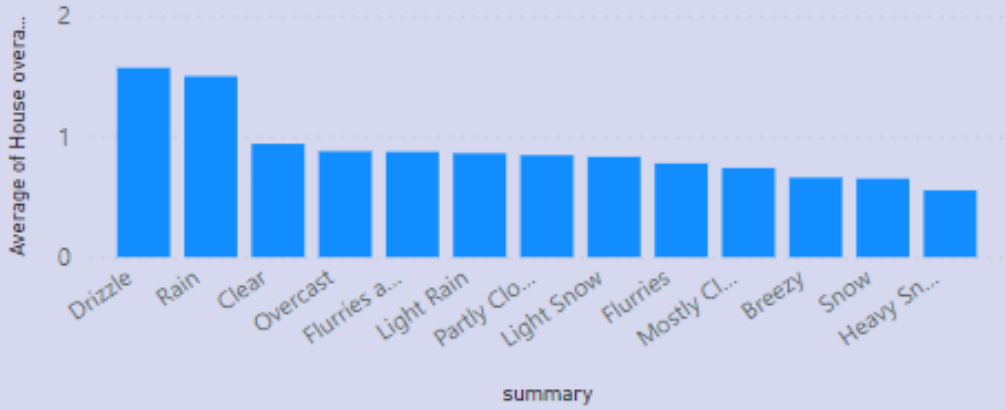


Cost of Electricity by Days

House overall [kW] and gen [kW] by Month



Average consumption [kW] in different weather conditions



1.98

Min of temperature

28.60

Average of temperature

Analysis of Dashboard: Custom Time Period (1/2)

- The dashboard summarizes the following analysis from the given data:
- We provide a custom time period from the total time covered in this dashboard, which then produces the visualization based on data retrieved from those dates/times.
- The dashboard tells us the electricity generation in kW during the time period as well as the cost in PKR during our provided time period.
- Compares the electricity generation and consumed during this period through gauges and a stacked bar chart so we see the numbers and its visualization at once.

Analysis of Dashboard: Custom Time Period (2/2)

- Describes the usage of electricity in rooms during this time period,
- Shows the average electricity consumption in different weather conditions during our chosen time period,
- Shows the total kilo-Watts electricity consumed in this time,
- Shows us the minimum temperature during this time as well as the average temperature during the duration of this time period.

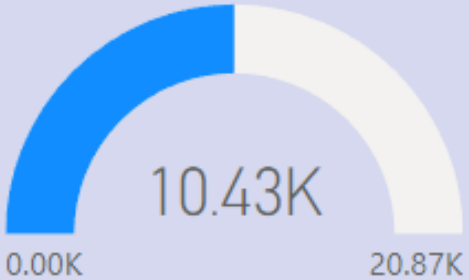
kW Electricity Consumed

432.01K

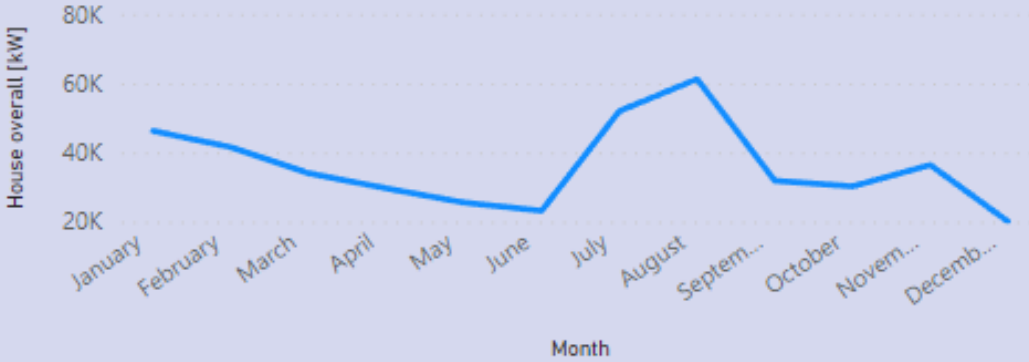
House overall [kW]

Cost spent on Electricty in PKR

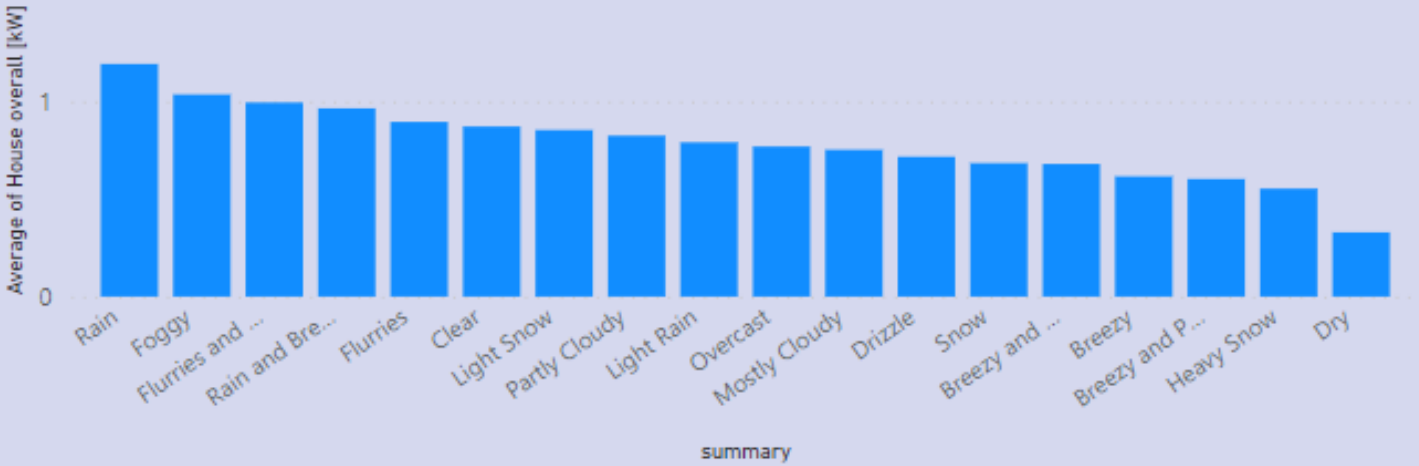
Cost saved on Electricity in PKR



House overall [kW] by Month

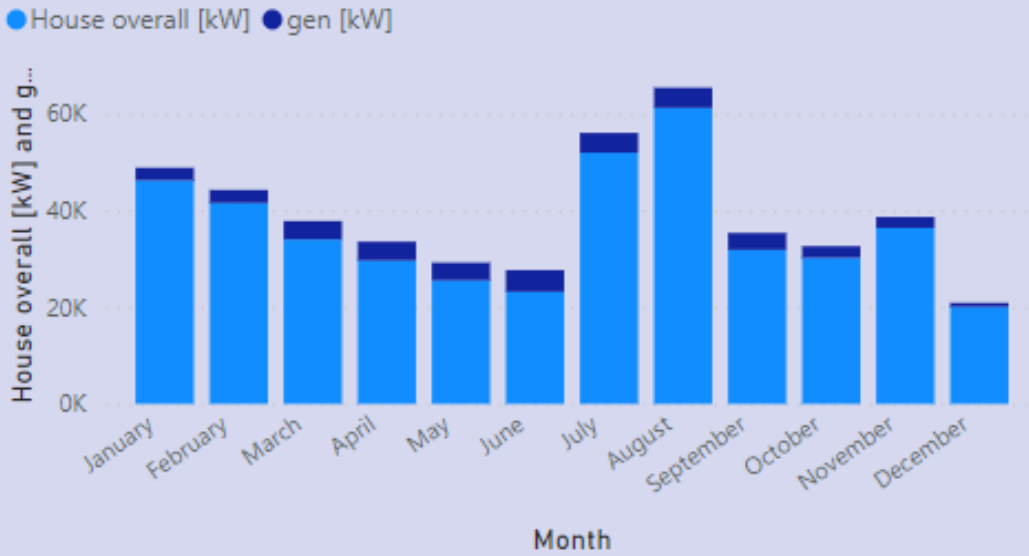


Average consumption [kW] in different weather conditions

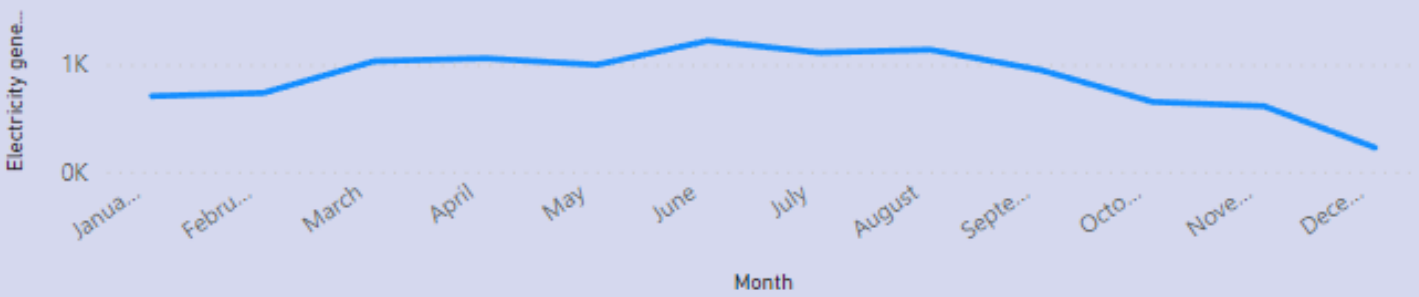


Cost Electricity by Month

House overall [kW] and gen [kW] by Month



Electricity generated by Month



Analysis of Dashboard: Electricity Utilization(1/2)

- The dashboard summarizes the following analysis from the given data:
- Demonstrates how the electricity is consumed in the house during different months.
- Demonstrates how much the electricity costs throughout the year in different months.
- The average consumption, in kilo-Watts, during different types of weather conditions such as rainy, dry, etc.
- Since we are also generating electricity by using solar power, it shows how much electricity we are able to generate using solar power.

Analysis of Dashboard: Electricity Utilization(2/2)

- Compares the generation and consumption and visualizes ot in the form of a stacked bar chart.
- Shows the cost spent, in PKR, on the consumption of electricity in the year 2016.
- Shows the cost, in PKR, saved as a result of electricity generation.
- We weren’t initially given the costs of electricity in the dataset. We have used the KE kW/minute rates in 2016 to calculate these measures and added it to our data.

KPIs:

1. Electricity Usage in the smart home in kilowatts,
2. Cost of electricity used IN PKR,
3. Electricity being generated through solar panels in kilowatts,
4. Cost saved by electricity generation in PKR
5. Factors increasing or reducing Consumption and electricity generation,

Final Analysis and Findings:

Story:

The three dashboard tell the following stories respectively:

1. Analysis during custom period:
   We see that it provides the key indicators during a specific period. In the picture, first half of December is selected, so we see that due to winters, the consumption is low especially during the latter days. Mostly the weather is cold (snowy and overcast) and the minimum and average temperatures are low too. The total electricity consumed in just almost 19000 kW, which is less than 10% of the consumption. Also shows the rooms most consuming the electricity during this period.

2. Overall Electricity Utilization:
   The third dashboard displays the following story, it shows that the electricity cost and generation was slowing down during the first months of 2016, which indicates a spring and an intermediate winter season. The cost and consumption went drastically up due to summers after June and it started coming down at a major rate after October. It also shows that during summers the electricity generated was the highest, that's because the generation method is solar energy. It also shows how much the consumption end generation were during different months. Then goes on to show the average consumptions during different weather conditions and we see that the weather condition mostly related to summers are on the right of the stack bar chart which means that during summer most of the electricity is consumed as indicated by the other visuals as well. Finally, we see the cost that is saved and the cost that is spent on electricity throughout the observation, and the number of units consumed.

3. Appliance/Room Consumption:
   Some of the most appealing visuals have been added to this dashboard. This dashboard first and foremost tells us the consumption of **all** the categories throughout different months of 2016. It then goes on to compare the electricity consumption of different appliances and shows that how these appliances are used the most in different months, and when they are used the least. The dashboard also shows that the break-up of the total consumption used by different rooms and by different appliances though tree map and pie chart respectively. Thus, we can see which appliances/rooms used the most and which appliance/room uses the least amount of electricity. Finally, we see the total number of KW consumed during the analysis. And we can derive the percentages through that number

Key Analysis:

We have figured out that the consumption of electricity throughout different months of 2016 rely heavily on the weather conditions. The peaks in graph show that electricity consumption has been significantly high during MAY-JULY which is the main time of summer. Along with this, the solar power generation that is being done to support the electricity consumption it is not enough as the readings taken in this data set are from a city which has a climate susceptible to rain throughout the year, thus the solar power generators does not get enough sunlight to match the needs of electricity consumption. We have also seen how different appliances affect the electricity consumption but more importantly if we do not consider the furnace used for a second, most consumption is saturated inside the Home Office, the living room which are the primarily used rooms/areas in a house.