Data Science @ Jazz- Preliminary Analysis

'Modelling Customer Recharge in Major Cities using Weather Information'

Danish Farid, August 2017 (**danishnxt@gmail** ***- in case any questions pop up)***

We began with the idea to build a model for predicting customer recharge behavior, based on expected weather conditions.

Let us first define "**weather**". Intuitively it makes sense for a model to factor in: Temperature, Humidity, Rain and Wind forecasts, but there is other information that we may not realize to have a noticeable impact on customer behavior: Sea Level Pressure, Dew Point Average, Visibility, etc.

About expected performance and value, the perfect model would be able to predict recharge numbers on a per 'cell-site' locality, but the ambition of the project is limited by what data we're able to reliably gather, now and in the future. Hence, at the start we weren't aware of what we could potentially end up with.

Data Sourcing:

The first step was to gather weather information, this was found online (wunderground.com), with historical monthly data pulled from airport weather probes for the three major cities, Lahore, Karachi and Islamabad/Rawalpindi. Airport ICAO codes were used to identify them easily on the website (OPLH, OPKC, OPRN), and all available data was copied into excel, cleaned and then saved as a TSV (Tab Separated Values) file for reading into Python. Hour-wise data was not considered at the time. Temperature data from Wunderground was crosschecked with data from Accuweather.com, for the last nine months (temperature Hi's and Lows) for the three cities. Both sources provided data that followed the same trends for temperature rise and fall, the small variation present was expected.

➔ Another source of data was learnt of later "http://ogimet.com/index.phtml.en", but was unused during this preliminary phase, it may be worth considering down the road.

Recharge Numbers were pulled via queries from the Jazz data warehouse. Daily recharge amounts and unique recharger numbers for 3 cities (Islamabad and Rawalpindi were added together as one city) were used.

Later during the project, we also required sunset and sunrise times for finding out day length variation. This data was taken from SunsetSunrise.com which uses a standard formula to generate sunset/sunrise times based on area Longitude and Latitude. This data was crosschecked from another popular website and appeared to be accurate. The trends they followed were also intuitively recognizable, and showed up on graphs, as they should have.

Following this, came visual data analysis in-line with phase 2 of the standard CRISP method. We plotted graphs on multiple time scales, of recharge information against different weather information to find a meaningful relation. Down the road, some degree of data normalization was required as well, detailed later.

Analysis: Technique and Results

Data was imported into Python (Anaconda, with the Spyder IDE) with the ‘**Pandas**’ library for data manipulation, ‘**MatPlotLib**’ for data plotting and ‘**SciKitLearn.Preprocessing**’ for data scaling. TSV files were imported using Pandas, labelled, filtered and (for weekly numbers) aggregated. To make aggregates easier, all daily records (recharge, weather, day length) were also tagged with a ‘week’ and ‘month’ field. Daily entries had a field to indicate what week/month they were part of in our total time-period [By the end, our full data range was from Aug 2016 to Aug 2017]. It allowed Pandas ‘groupby’ function to be extremely useful, making weekly aggregation a breeze.

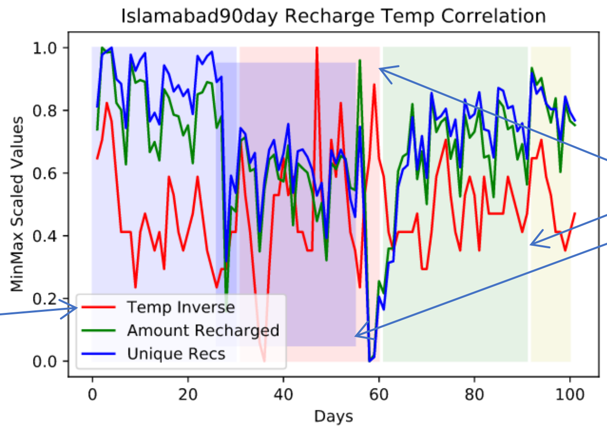
Additionally, all data was scaled to a 0-1 range for normalization, with every field scaled independently, using SciKitLearn MinMaxScaler objects. The numerical information lost didn’t matter since we merely wanted to visually identify presence of a relationship.

Upon initial plotting of the past 4-month data, some things became immediately clear:

The graph was plotted on a ‘per day’ scale, so there are many factors at play here. Normal monthly and weekly customer trends both made looking for the effects from weather harder. We realized that daily data had too much noise to be able to confidently visually identify meaningful correlation.

Temperature data was inversed, post-scaling, to allow increase to mean “better weather”

Now ideally, this should spike **with** recharge information



The backgrounds represent the different months, May to start of August

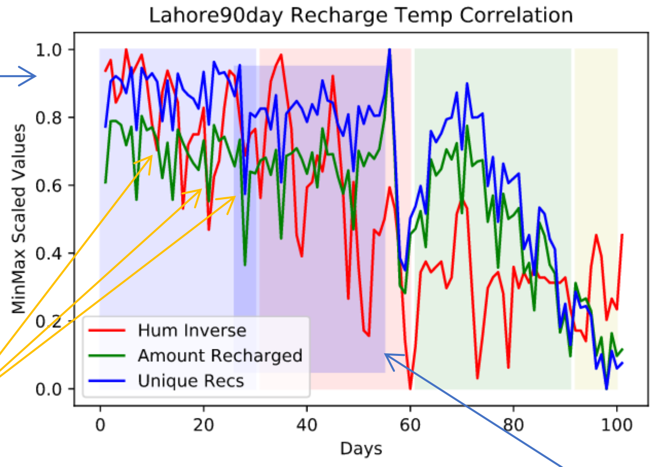
The smaller box in the middle represents Ramzan

But while a sure-fire correlation wasn’t possible, we could see trends in some areas. Additionally, our confidence was further boosted by observation of the Recharge/Humidity Graphs.

Due to the different scaling factors, the size of dips and spikes will be different. But we were only looking for them to match up, no matter their size

Notice in some areas that recharge and humidity (inverse) spike together, indicating possible relation.

Due to the noise though, still difficult to confirm exactly

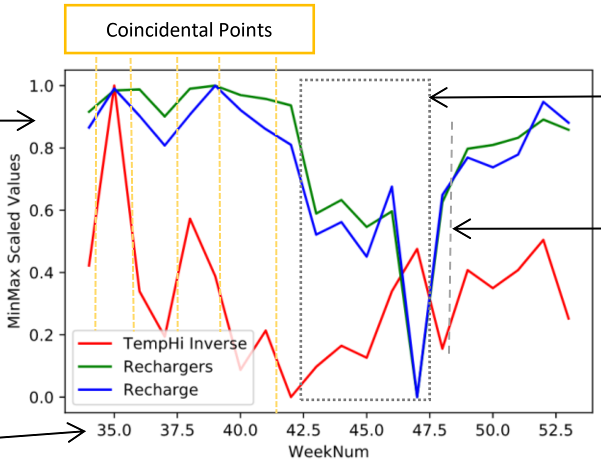


Ramzan and Eid both come with their own consumer behavior trends as can be seen here.

The characteristic rise and fall pre/post Eid has less do with the weather, but since the weather happened to coincide with it, in a way we hoped, it may mess up an accurate model later.

This was something to consider

We hypothesized that this stronger link between recharge and humidity may be because humidity has more of an impact on the consumer behavior by more directly affecting what the day “feels like” than temperature. Moving forward though, it was clear that we had to move to a different time scale, and that weekly averages were the way to go to visually identify aggregated trends better without daily noise. Monthly trends would still exist, but then they would be consistent, month on month (except during Ramzan). The weekly graphs are as follows.



Islamabad 20 Week TempHi Correlation

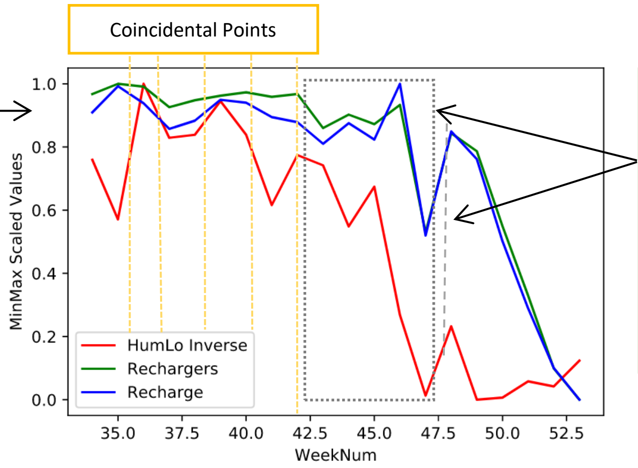
Again, look only for coincidental spikes, disregard size

WEEK NUMBERS IN GRAPHS ONWARDS START FROM WEEK 0 (AUG’ 16) up to WEEK 53 (AUG’ 17)

Ramzan Still Skews the graph overall, reducing usability for normal day trends

The climb back to normal (Post Eid) again coincides with better weather, even though it probably has less to do with it, the same problem as the previous graphs

Temperature shows some promise before Ramzan, as well as after. But after appears to be more coincidental with the slow recharge rise after Eid, this can lead to the false hope of an airtight model. Let us now look the recharge numbers against humidity for same time period.



Lahore 20 Week HumLo Correlation

You’ll notice that the humidity spikes here are odd, they don’t line up with temperature spikes.

This prompts further research into exactly how the two interact with each other, more on this on the next page.

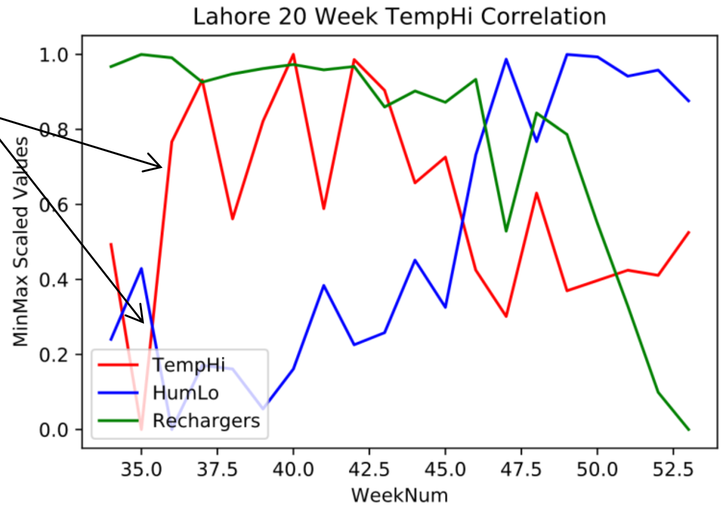
The Ramzan and post Ramzan period was worth removing from our analysis from here on out, to consider normal days only.

During model generation, this is something that’ll also have to be considered.

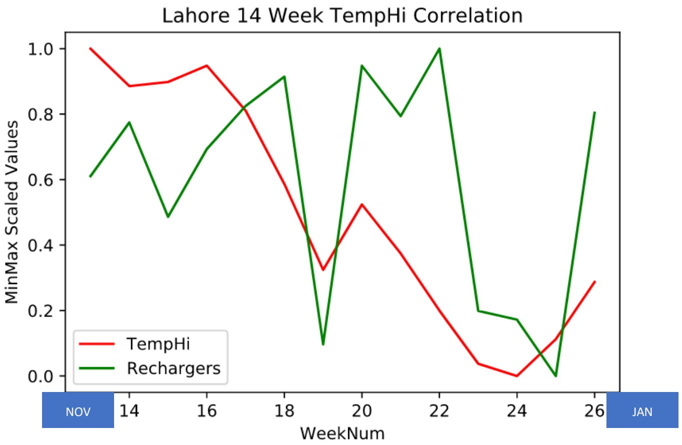
Following are a few more graphs. The first is a combination of the two above, with both Temp and Humidity against daily recharges to compare their trends and see their relationship.

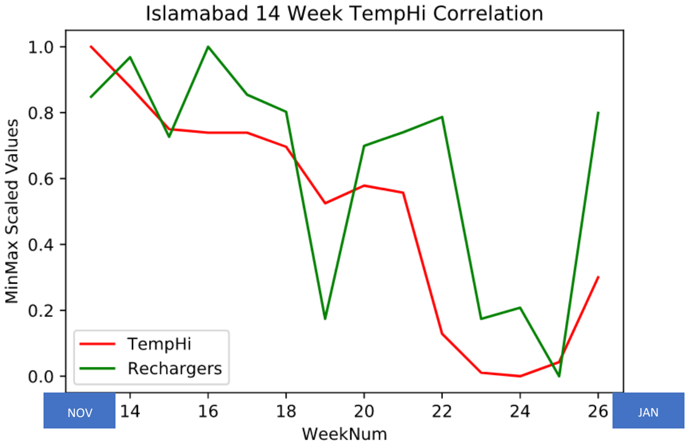
The two after are from the **Winter Period in Lahore and Islamabad**.

Notice the inverse relationship between Humidity and Temperature



This relationship prompts a deeper consideration into how the unique features of weather interact with each other, but the obvious is visible. The higher the temp, the less the humidity on our scaled range. This will be important.





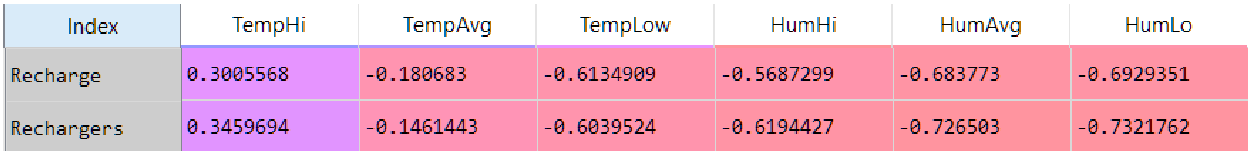
Both graphs above cover the November, December, January Period.

So far, we've seen only Lahore and Islamabad. The same trends are present, but less pronounced in Karachi. Expected **Winter trends** show up better than expected **summer trends**. This is probably because in Karachi, the "feels like" weather varies differently from that in land locked cities like Islamabad and Lahore. July for instance, in Karachi is supposed to be better than June, with cooler winds blowing from the ocean. This indicates that we may need different models for distinct types of cities.

Analysis Phase: Moving to the Numbers

During this preliminary walkthrough, I will use one city at a time to show you how we developed our understanding of what was happening with the numbers and then present 6 matrices for the 3 cities in 2 seasons (hot and cold) at the end.

Calculating for Lahore in the April to August 2017 period led to the following matrix.



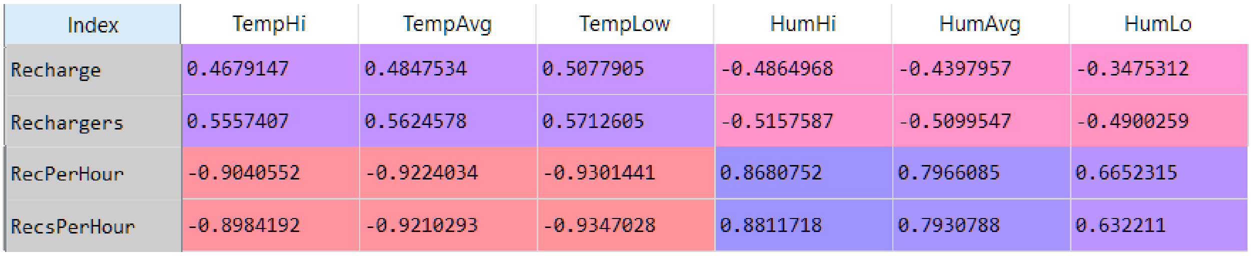
Index	TempHi	TempAvg	TempLow	HumHi	HumAvg	HumLo
Recharge	0.3005568	-0.180683	-0.6134909	-0.5687299	-0.683773	-0.6929351
Rechargers	0.3459694	-0.1461443	-0.6039524	-0.6194427	-0.726503	-0.7321762

Correlation Matrix . Lahore . April to August 2017

Humidity and Recharge/s numbers show up as negatively correlated, this is what we hoped for. However, the 'TempHi' numbers and 'TempAvg' numbers show the wrong, and unexpectedly weak correlation respectively.

In addition, we saw visually how temperature and humidity are inversely related, but TempLow appears to be in line with humidity here. This discrepancy was confusing as to how one of the temp field acts differently from the others. We believe this is a combination of two factors: The skew and upset of trends caused by Ramzan, and another intrinsic factor hidden in temperature information that we didn't realize: Length of the day. Which has (through daylight duration) a strong link with recharge behavior.

To improve on this, we took two steps. We updated the observation period now starting at a point when the weather starts to get hotter after winter, to just before Ramzan. From Feb to end of May. Also, we normalized the daily recharge numbers with the day length information (effective daylight) to get recharge/s per hour, hopefully reducing the effect of the day length variation on recharge. Doing this resulted in the following:



Index	TempHi	TempAvg	TempLow	HumHi	HumAvg	HumLo
Recharge	0.4679147	0.4847534	0.5077905	-0.4864968	-0.4397957	-0.3475312
Rechargers	0.5557407	0.5624578	0.5712605	-0.5157587	-0.5099547	-0.4900259
RecPerHour	-0.9040552	-0.9224034	-0.9301441	0.8680752	0.7966085	0.6652315
RecsPerHour	-0.8984192	-0.9210293	-0.9347028	0.8811718	0.7930788	0.632211

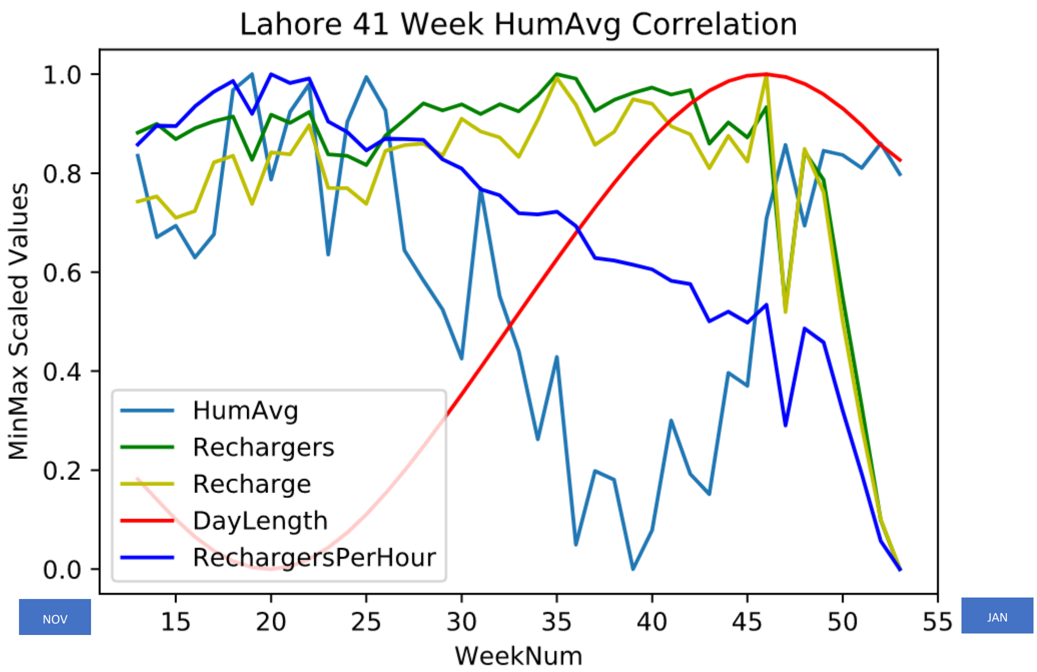
Correlation Matrix . Lahore . Feb to Mid-May 2017

This has mixed results. As we have removed the most humid parts of the summer from the observation period, humidity correlation is brought down slightly. But, the Recharge/s per hour (normalized fields) are now in line with what we expected them to be against temperature. However, in the effort to normalize for day length, humidity appears to have reversed its effect entirely, when compared against recharge/s per hour.

While this may appear to be a very random result, bear in mind how we confirmed visually that humidity and temp are inversely linked. Also, the most humid parts of the year **have** been removed to take Ramzan out of the observation period. Additionally, after some discussion we believe that this brought forth another interesting result. Temperature/day-length appears to affect consumer recharge behavior during the day, **more** than it affects how much or how many people recharge that day.

To see just how much of an effect daylight change has on recharge, following is a plot from Nov'16 to Aug'17 to see it's long term effects. With it, we hope to find an answer to the following question: **Does Jazz actually get less loading on shorter days?**

We can use this graph to compare seasonal variation as well as the difference across a season. You will notice in the Summer (which starts about midway), that while the days get longer and the raw recharge numbers have an uptick, it is not as drastic as the downtick of the daily recharge/hour density. Meaning, people will now recharge across a wider range of time than they will recharge more. This confirms our hunch that recharge behavior is changing more than raw recharge numbers. But, both are taking place simultaneously as we can see. The raw recharge uptick will also be because of the usual seasonal trends.

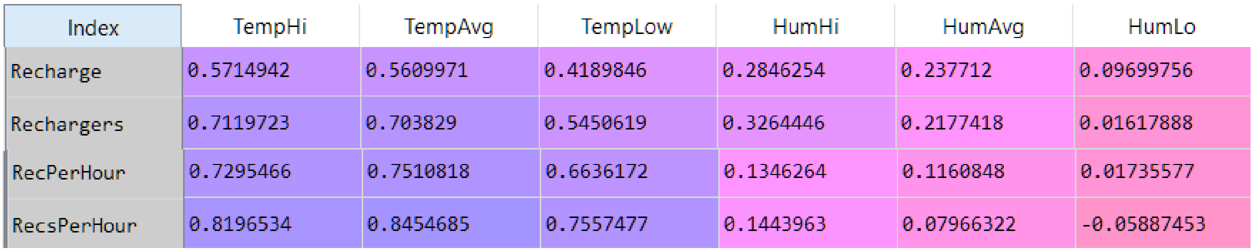


We also see here that the humidity falls initially with the recharge per hour. This is during a period when humidity isn't extremely important. But as far as the correlation coefficients go, it skews the values and causes strong positive correlation as they both decrease gradually together.

Humidity we've seen to have an effect, but mostly within the three critical months of the year: June, July, and August, especially in places like Lahore and Islamabad. Unfortunately, due to Ramzan overlap, we had to remove those periods from our analysis. So, we're not able to factor that in.

Additionally, day length will increase regardless of how humid it is, early summer. And it is only during the later months when the effective daylight fluctuation slows down, that the daily humidity and temp fluctuations, will affect recharge more. Keeping these two phenomena apart proves to be tricky, we must pick and choose our observation period to feed the model for a certain time of year very carefully. If humidity really kicks in for just 3 months in the year, looking at past year data in where Ramzan overlap is not present during those months will help form a better relationship with it.

Let us now look at the opposite, the winter period. For sake of simplicity we'll look at the peak winter months, **December and January,** in **Karachi** and see what clues we can pull from it.



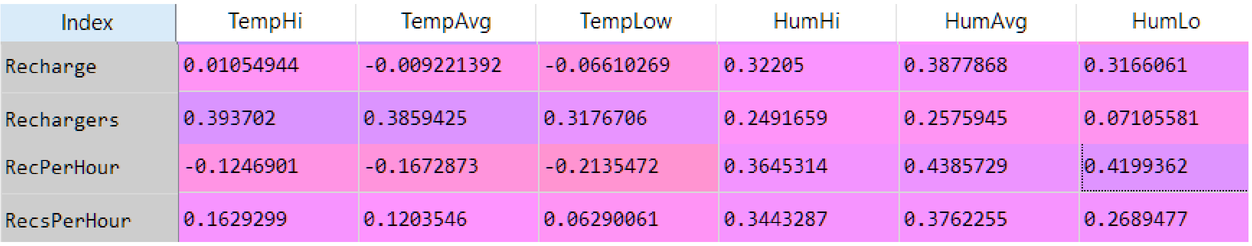
Index	TempHi	TempAvg	TempLow	HumHi	HumAvg	HumLo
Recharge	0.5714942	0.5609971	0.4189846	0.2846254	0.237712	0.09699756
Rechargers	0.7119723	0.703829	0.5450619	0.3264446	0.2177418	0.01617888
RecPerHour	0.7295466	0.7510818	0.6636172	0.1346264	0.1160848	0.01735577
RecsPerHour	0.8196534	0.8454685	0.7557477	0.1443963	0.07966322	-0.05887453

Correlation Matrix – Karachi – Dec'16 to Jan'17

These are encouraging results, we expected High Temp to correlate with High Recharge. It appears to have some degree of an effect on the raw numbers as well as the normalized recharge density numbers.

Not only do we get more recharge on a day, it's also denser. We can expect this since in the winter, a good day will feel better during certain periods, not the entire duration. It is likely that people will again cluster up to recharge in certain "good" hours, as well as possibly recharge more, this bodes well for the model.

To extend this, lets incorporate November 2016 as well to have a longer, less 'extreme-biased' observation period and see what happens to the numbers.



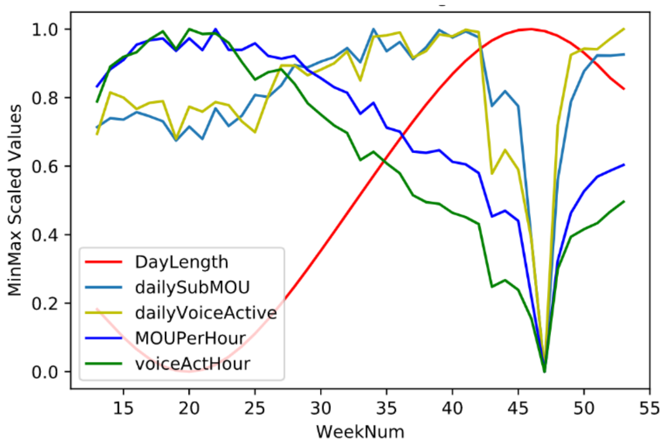
Index	TempHi	TempAvg	TempLow	HumHi	HumAvg	HumLo
Recharge	0.01054944	-0.009221392	-0.06610269	0.32205	0.3877868	0.3166061
Rechargers	0.393702	0.3859425	0.3176706	0.2491659	0.2575945	0.07105581
RecPerHour	-0.1246901	-0.1672873	-0.2135472	0.3645314	0.4385729	0.4199362
RecsPerHour	0.1629299	0.1203546	0.06290061	0.3443287	0.3762255	0.2689477

Correlation Matrix – Karachi – Nov'16 to Jan'17

The numbers fall. Correlation is weaker across the board since November is a less extreme month. There is however still some temp correlation with daily Rechargers. But this, again, indicates the importance of picking the observation periods for the data properly. If you run an analysis on the extreme months, the data is richer and the results more accurate. If we extend that period to data that is less rich and precise, the results will become less accurate.

Finally, after seeing customer recharge behavior change with the features of the day, we decided to look at how other types of behavior change as well. Customer MOU and Daily Voice Active numbers for the last 12 months were graphed with the day length information. This was useful, but it was only when we normalized these numbers with the length of day (much like recharge information) that the picture became much clearer.

The graph below is for Islamabad during an Observation period from Nov'16 to Aug'17. Ignore the extreme dip in the post EID period. This is a side effect of the scaling, we can safely ignore it.



This confirms what the previous graph showed in terms of customer recharge behavior. Daily MOU and Voice Active Numbers go up in the summer, but not as fast as the day length increases. So, we **are** getting seasonal variation with recharge as well as revenue generating activity. However, it is not proportional to the day length variation. Meaning, people will now call and generate similar revenue in more varied parts of the day, rather than generate more of it during the day.

Summary

While we started with just hoping to lock in a relationship between raw recharge numbers and basic weather factors such as temperature and humidity, we ended up discovering much more at play. Both in the weather domain and the customer behavior trends, for recharge as well as revenue generating behavior.

Following are some of the findings we can walk away with.

- ➔ There is a definite link between the weather and consumer behavior, although we may have to do deeper, to a finer resolution than just daily trends to be able to make the best use of it.
- ➔ The link between weather and consumer behavior is prevalent in both seasons, with their own nuances.
- ➔ A hidden factor within temperature for the length of day has an enormous impact, independently from how the day feels, on consumer behavior.
- ➔ Daily consumer behavior may be more impacted than the raw recharge totals moving from season to season, and month to month.
- ➔ The weather has their own links and relations. Temperature and humidity are inversely linked, for instance, which we must incorporate by looking at previous year data perhaps.
- ➔ Keeping apart direct factors of temperature (feel of the day) from indirect factors (length of the day), requires careful selection of our observation periods and removal of outliers/normalization to eliminate seasonal variation.

To end, I will expand on the last finding, please refer again to the last graph. For complete dependence on day length, the daily Voice Active and MOU numbers would have had to increase in proportion to the increase in day length in the summer. This would have kept MOU/VoiceActive per hour densities the same, instead of decreasing. Clearly this is not the case, and I believe people will naturally adapt their habits to the day as it grows and shrinks and feels better or worse. But are less likely to forgo certain habits entirely on a shorter or less pleasant day.

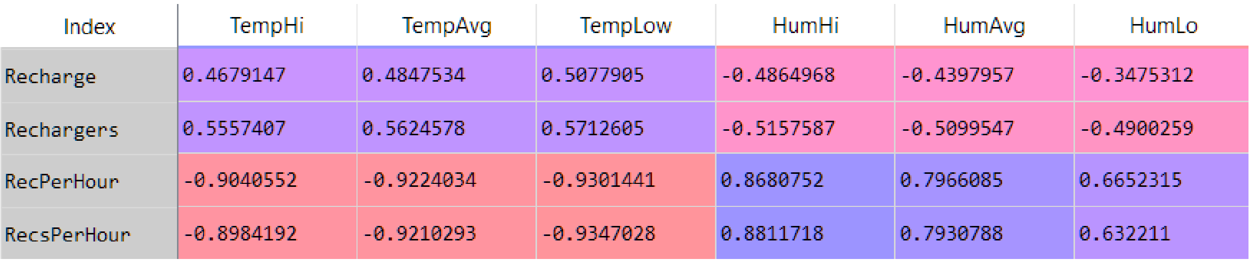
The trick is to find people **who** will (forgo certain habits), and the specific period of the year **when** they will do so. We will then hopefully be able to explain **why,** to benefit the business side of the organization. Which is the end goal of our analysis.

This is where model building usually leads into segmentation and further data normalization. But I hope that this short analysis will provide some insight into the domain and prove to be useful in some way. As already mentioned, if there are any questions at all, my email is at the top of the document. Feel free to reach out any time at all.

Appendix: Correlation Matrices for 3 cities, Summer and Winter (period defined below)

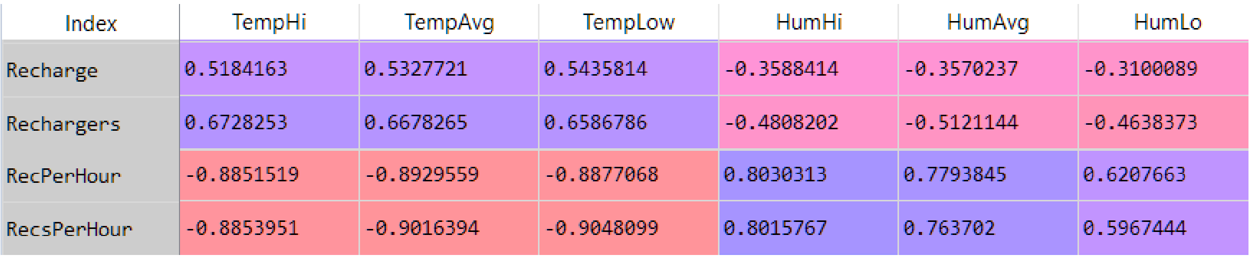
Summer [Feb – Mid May]

Lahore:



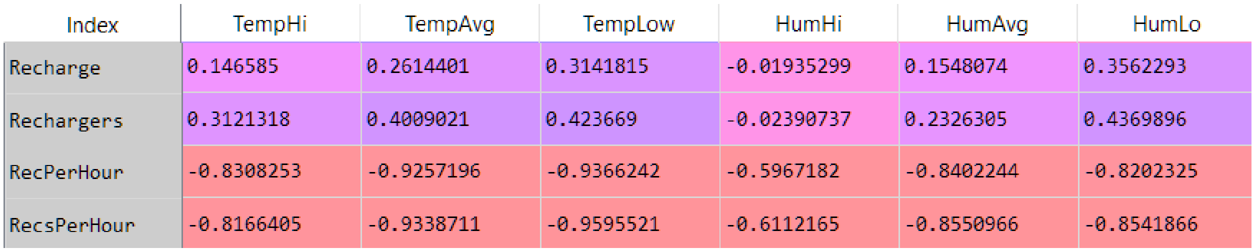
Index	TempHi	TempAvg	TempLow	HumHi	HumAvg	HumLo
Recharge	0.4679147	0.4847534	0.5077905	-0.4864968	-0.4397957	-0.3475312
Rechargers	0.5557407	0.5624578	0.5712605	-0.5157587	-0.5099547	-0.4900259
RecPerHour	-0.9040552	-0.9224034	-0.9301441	0.8680752	0.7966085	0.6652315
RecsPerHour	-0.8984192	-0.9210293	-0.9347028	0.8811718	0.7930788	0.632211

Islamabad:



Index	TempHi	TempAvg	TempLow	HumHi	HumAvg	HumLo
Recharge	0.5184163	0.5327721	0.5435814	-0.3588414	-0.3570237	-0.3100089
Rechargers	0.6728253	0.6678265	0.6586786	-0.4808202	-0.5121144	-0.4638373
RecPerHour	-0.8851519	-0.8929559	-0.8877068	0.8030313	0.7793845	0.6207663
RecsPerHour	-0.8853951	-0.9016394	-0.9048099	0.8015767	0.763702	0.5967444

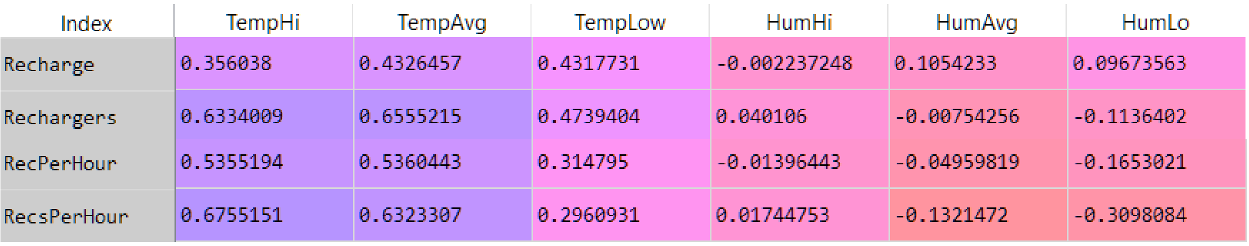
Karachi:



Index	TempHi	TempAvg	TempLow	HumHi	HumAvg	HumLo
Recharge	0.146585	0.2614401	0.3141815	-0.01935299	0.1548074	0.3562293
Rechargers	0.3121318	0.4009021	0.423669	-0.02390737	0.2326305	0.4369896
RecPerHour	-0.8308253	-0.9257196	-0.9366242	-0.5967182	-0.8402244	-0.8202325
RecsPerHour	-0.8166405	-0.9338711	-0.9595521	-0.6112165	-0.8550966	-0.8541866

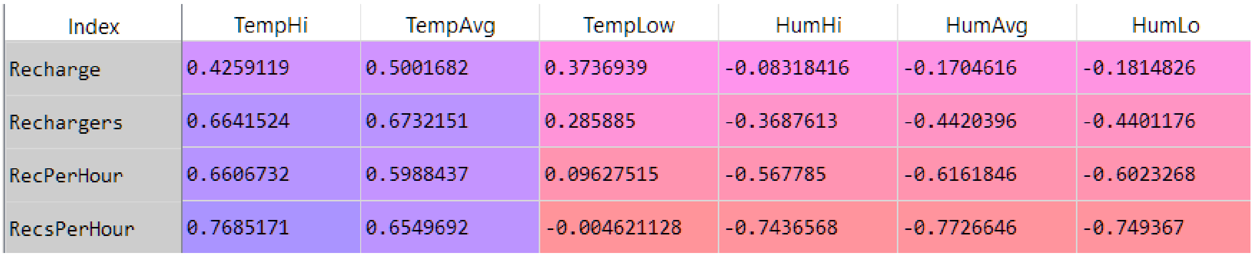
Winter [December, January]

Lahore:



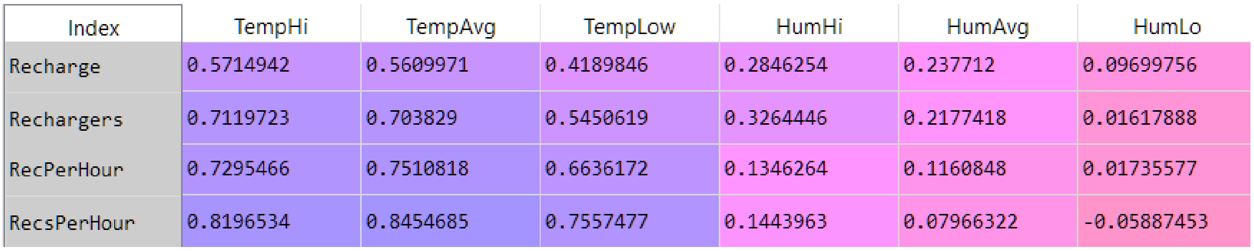
Index	TempHi	TempAvg	TempLow	HumHi	HumAvg	HumLo
Recharge	0.356038	0.4326457	0.4317731	-0.002237248	0.1054233	0.09673563
Rechargers	0.6334009	0.6555215	0.4739404	0.040106	-0.00754256	-0.1136402
RecPerHour	0.5355194	0.5360443	0.314795	-0.01396443	-0.04959819	-0.1653021
RecsPerHour	0.6755151	0.6323307	0.2960931	0.01744753	-0.1321472	-0.3098084

Islamabad:



Index	TempHi	TempAvg	TempLow	HumHi	HumAvg	HumLo
Recharge	0.4259119	0.5001682	0.3736939	-0.08318416	-0.1704616	-0.1814826
Rechargers	0.6641524	0.6732151	0.285885	-0.3687613	-0.4420396	-0.4401176
RecPerHour	0.6606732	0.5988437	0.09627515	-0.567785	-0.6161846	-0.6023268
RecsPerHour	0.7685171	0.6549692	-0.004621128	-0.7436568	-0.7726646	-0.749367

Karachi:



Index	TempHi	TempAvg	TempLow	HumHi	HumAvg	HumLo
Recharge	0.5714942	0.5609971	0.4189846	0.2846254	0.237712	0.09699756
Rechargers	0.7119723	0.703829	0.5450619	0.3264446	0.2177418	0.01617888
RecPerHour	0.7295466	0.7510818	0.6636172	0.1346264	0.1160848	0.01735577
RecsPerHour	0.8196534	0.8454685	0.7557477	0.1443963	0.07966322	-0.05887453