AI4C - Data Analysis

Henokh Y. Fibrianto

1. Understanding the data components

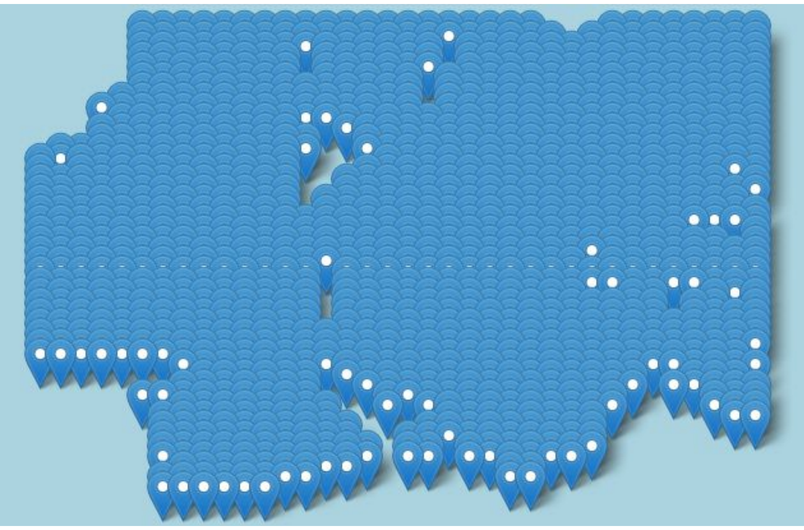
In the raw data (training.csv), there are four components, namely: geohash6, day, time and demand.

What are the meaning of each component in the data?

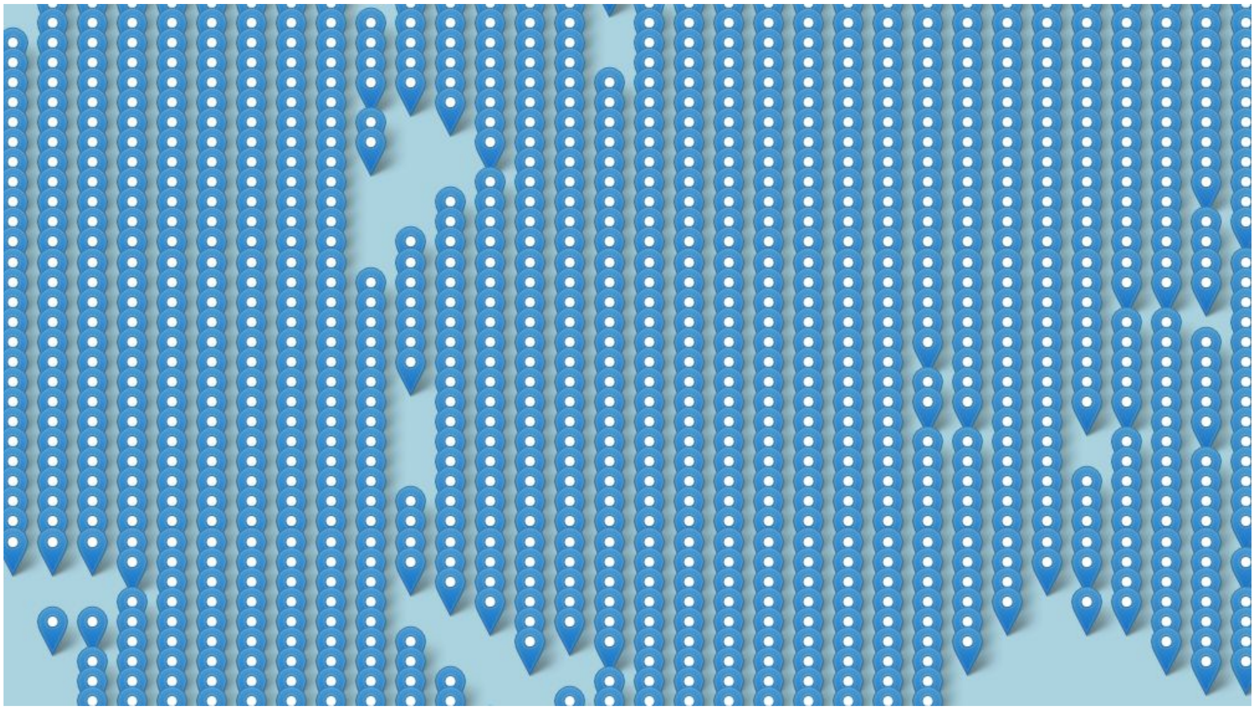
a. geohash6 : string representing hashed location information. Can be used as an identity key for each location.
b. day : Integer value starting from 1 to 61. Assumed to be sequentially correct and no missing day, therefore, it can be used to extract weekday information (monday, …, sunday).
c. timestamp : string representing time information with 'HH:mm' format and fixed interval of 15 minutes. Therefore, there are 96 unique values. Assumed to be sequentially correct.
d. demand : float ranging from slightly larger than 0.0 to 1.0 representing the normalized demand. Assumed to be normalized with respect to all demands regardless of location and time.

2. Understanding the geohash6

What can the decoded values of geohash6 tell us about geographical locations?
The decoded geohash6 data tells us about the location where the demand emerges as shown in the figure below.



Let's take a closer look.



Thanks to the limited precision of geohash, most of the locations (as represented by blue location pointer) has a fixed distance with the neighboring locations. Therefore, we don't have to do preprocessing to the location data (clustering, etc.), and each geohash6 value is a unique location.

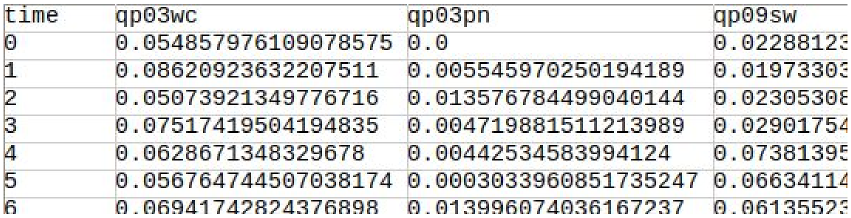
3. Understanding the demand

The demand can be understood by building a geo-temporal database containing the demand at each location (geohash6) during each time (day and timestamp). The columns of the database will be a list of the location and the rows will be the unique day and timestamp. Since we have 60 unique days and 96 unique timestamps, there are 60 x 96 = 5760 possible rows. The day and timestamp is converted into an integer of time-key using the following function.

Time-key = (day - 1) * 96 + timestamp.

Where the timestamp is the integer value of the time with 96 unique values, for instance 00:30 and 02:00 becomes 2 (2 * 15 minutes) and 8 (8 * 15 minutes), respectively. And the time key for the last time in the last day is (60 - 1) * 95 + 96 = 5759.

Finally, the database looks like the figure below.

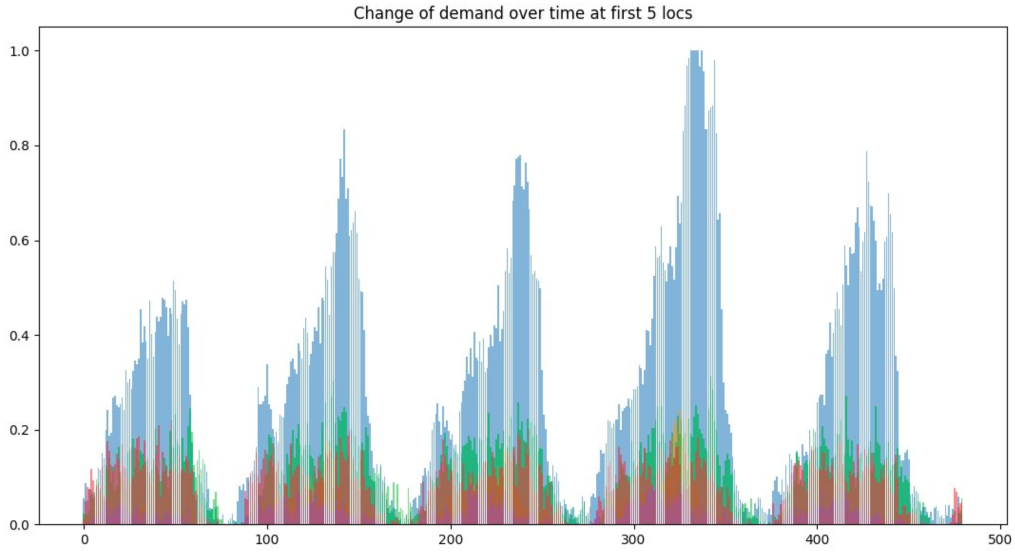


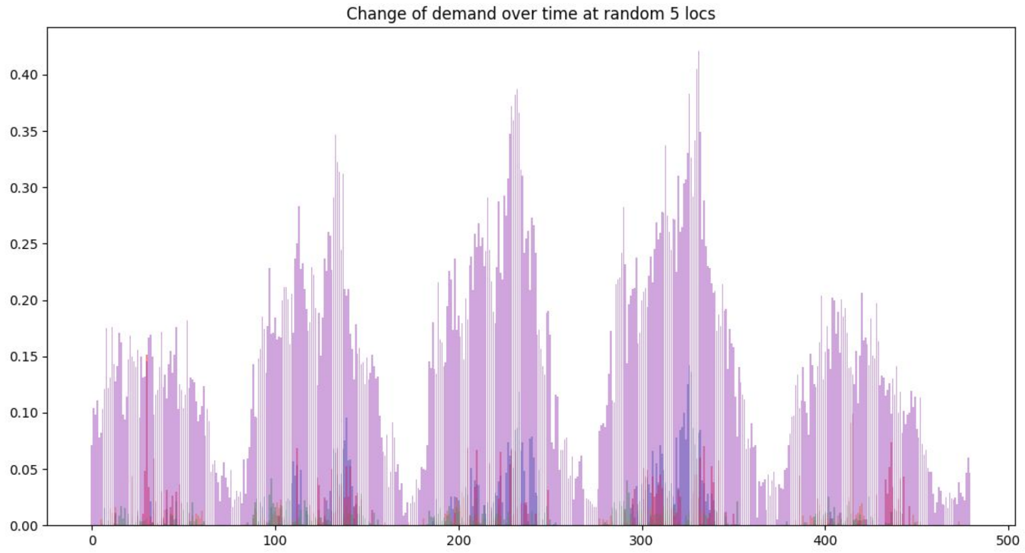
time	qp03wc	qp03pn	qp09sw
0	0.054857976109078575	0.0	0.02288123
1	0.08620923632207511	0.005545970250194189	0.01973303
2	0.05073921349776716	0.013576784499040144	0.02305308
3	0.07517419504194835	0.004719881511213989	0.02901754
4	0.0628671348329678	0.00442534583994124	0.07381395
5	0.056764744507038174	0.0003033960851735247	0.06634114
6	0.06941742824376898	0.013996074036167237	0.06135523

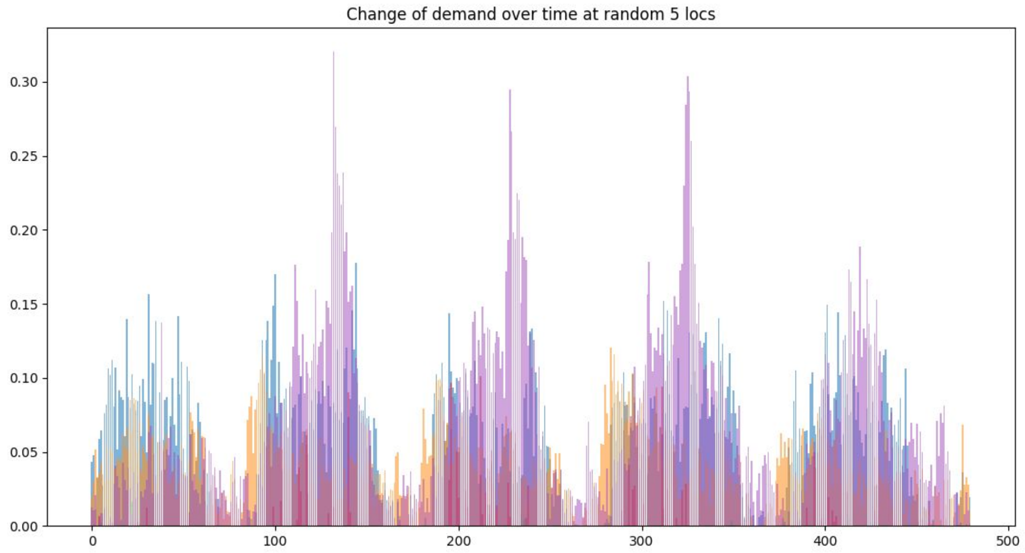
Please note that the figure only shows a tiny part of the database. In total there are > 1000 columns and 5760 rows. Once we have this database we can do more analysis about the demand.

How the demand looks like over time?

Below, we can see how the demand looks like at random locations in three overlapping bar charts, each containing the demand information of 5 locations (each location has different color within the same graph) for the first 5 days.



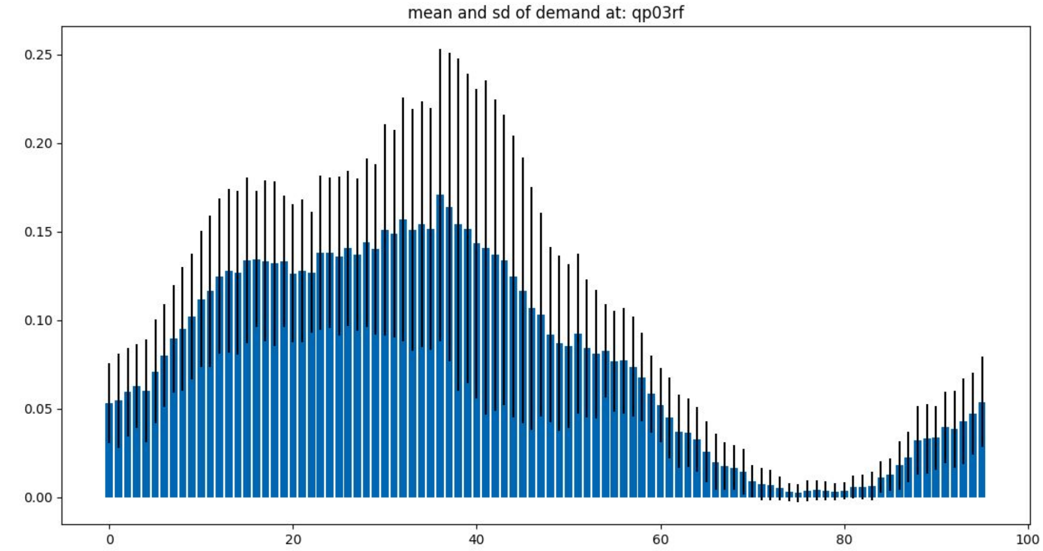


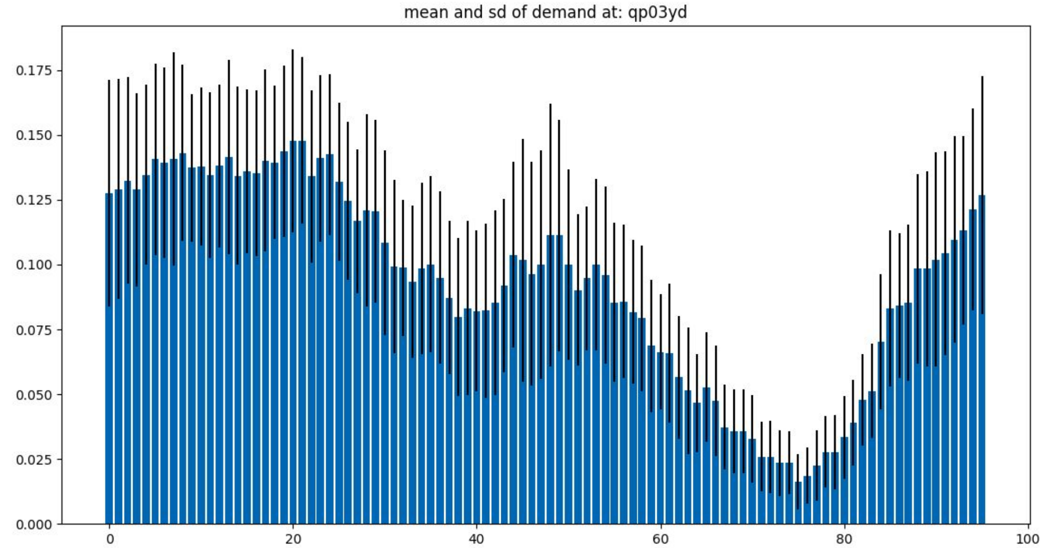


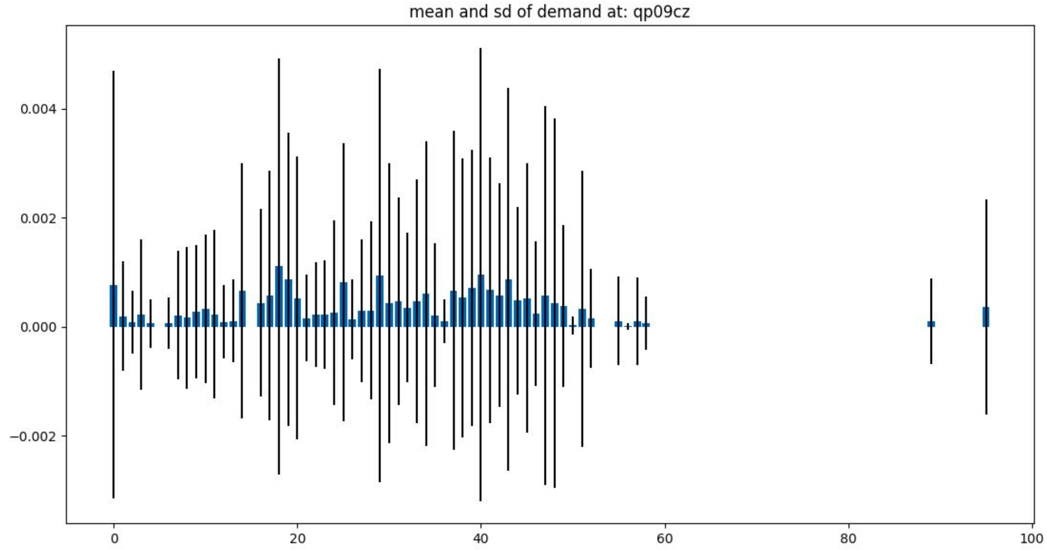
Notice that the demands at each location has a different pattern than the other location. Therefore, we have to characterize each location.

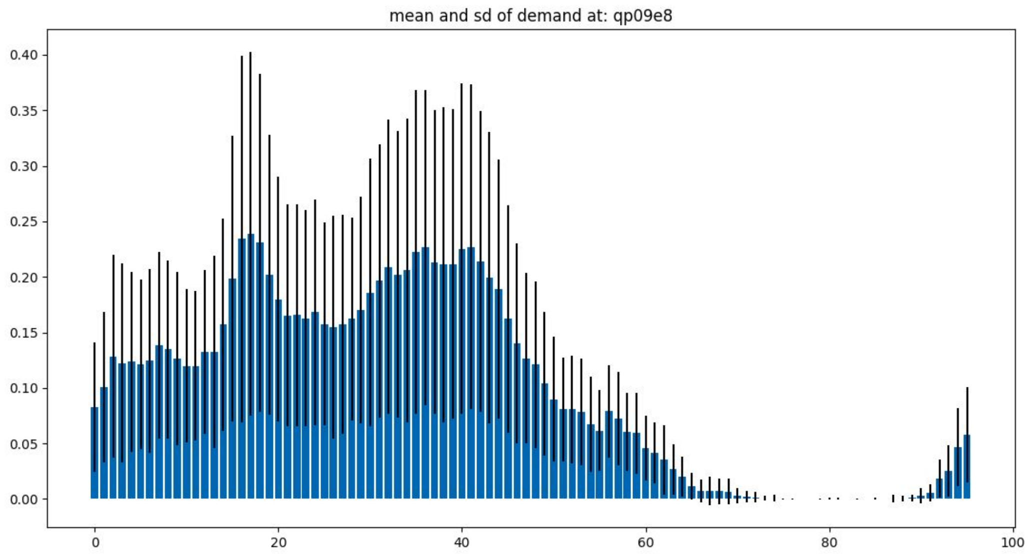
How can we characterize each location?

Since the pattern appears to be cyclic in daily basis, we can take the mean and standard deviation (sd) of the demand at each unique time in a day for each location and construct and 'demand profile' for each location, where the demand profile represents the demand pattern at the location. Below are the bar charts for a few locations with the demand as the height of the bar and the black line on each bar represents the standard deviation.


mean and sd of demand at: qp03rf


mean and sd of demand at: qp03yd

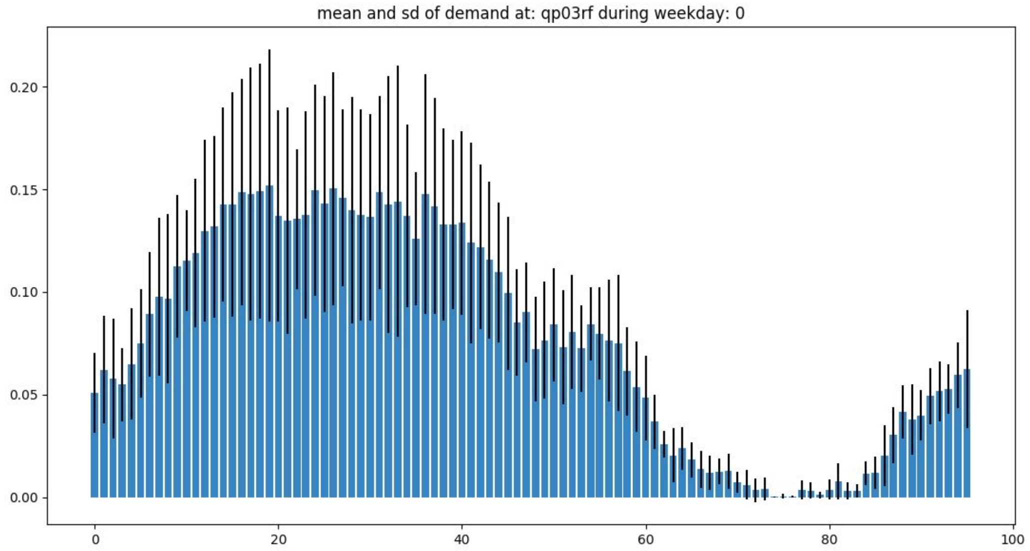

mean and sd of demand at: qp09cz

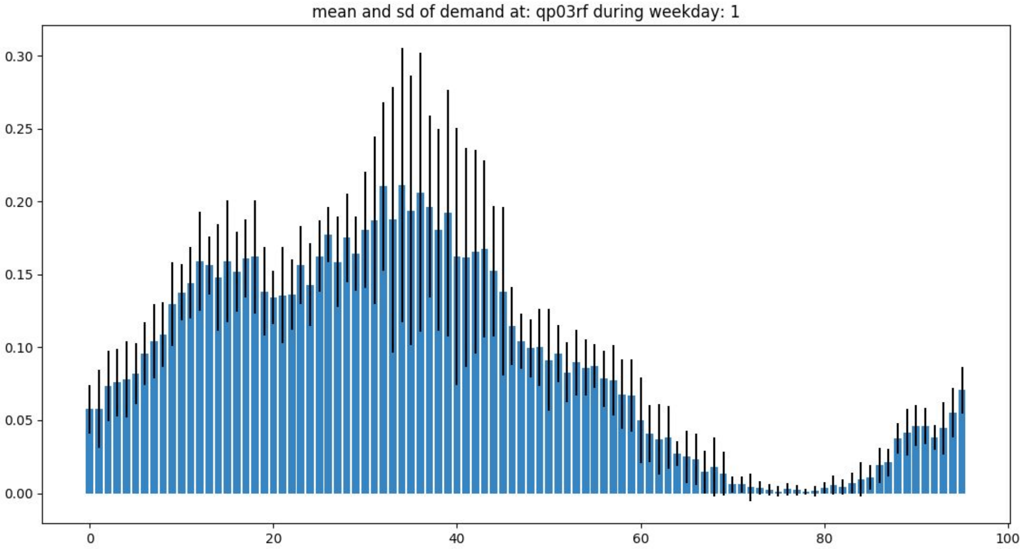


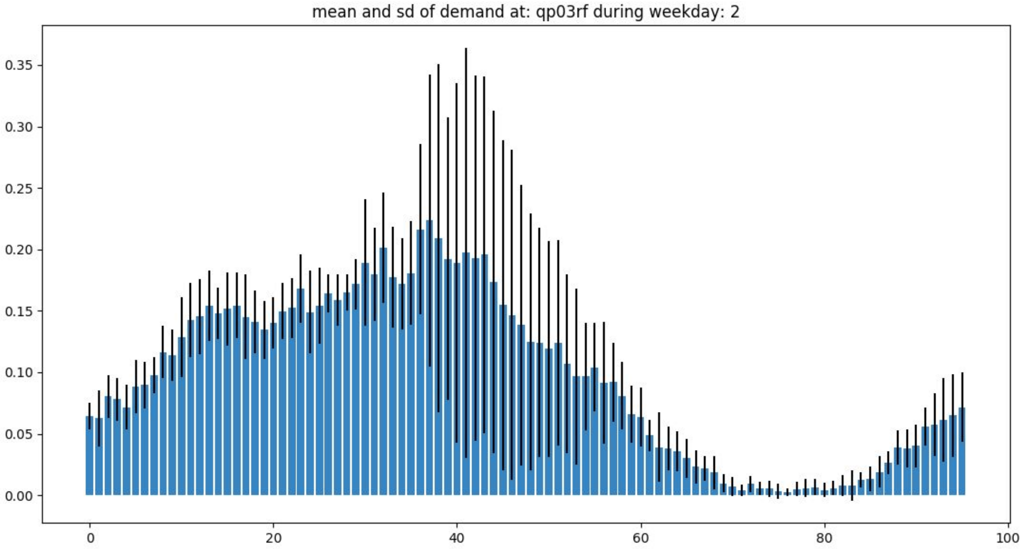
Based on the above figures, we can see how unique the demand pattern at each location is. However, there are still at least two information that can be discovered about the demand data, which are: does ‘weekday’ (names of days in a week: monday, ... , sunday) influences the demand pattern? and is there any pattern in the change of demand from one moment to the next?

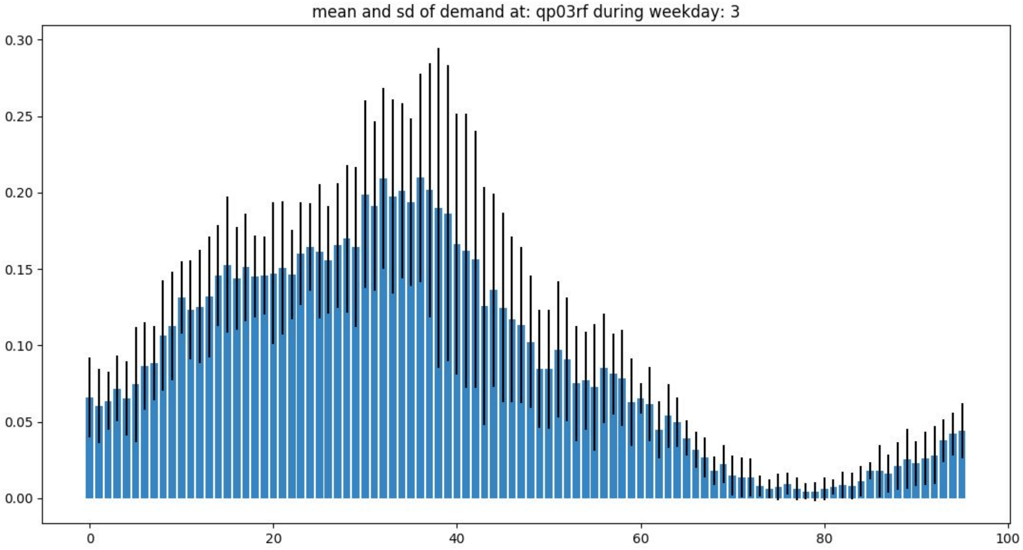
Does ‘weekday’ influences the demand pattern?

This question sounds sensible as people might spend more time travelling during weekend/weekday than any other days. Since it is assumed that there is no missing days and there are 7 days in a week in the data, we can extract the weekday information. The figures below show the demand at a location from weekday 0 to weekday 6 at the same location.

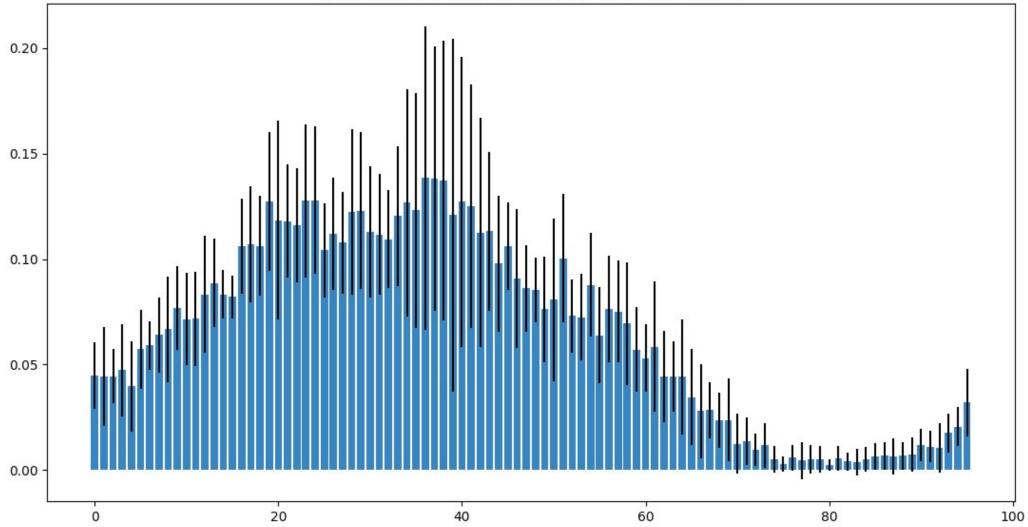



mean and sd of demand at: qp03rf during weekday: 1
0.30
0.25
0.20
0.15
0.10
0.05
0.00
0
20
40
60
80
100

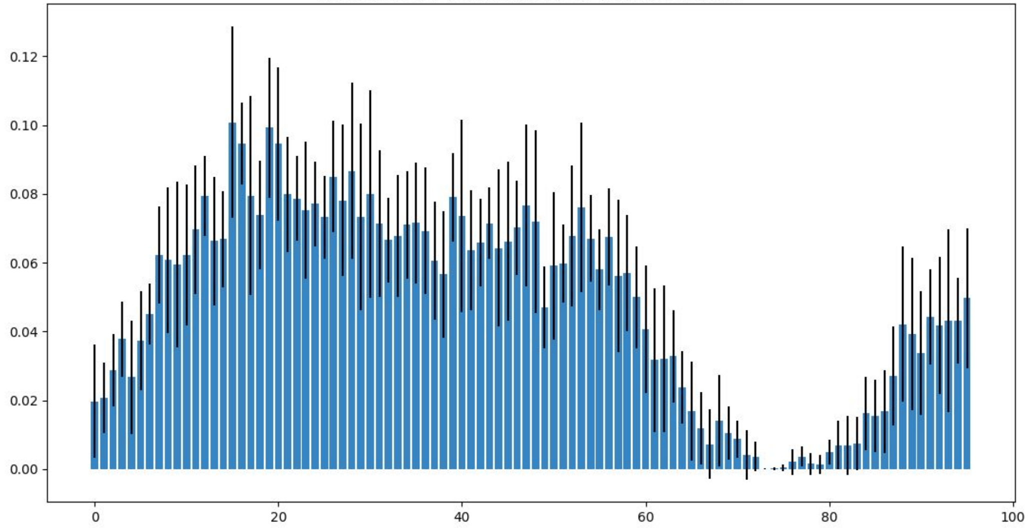

mean and sd of demand at: qp03rf during weekday: 2
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
0
20
40
60
80
100


mean and sd of demand at: qp03rf during weekday: 3
0.30
0.25
0.20
0.15
0.10
0.05
0.00
0
20
40
60
80
100

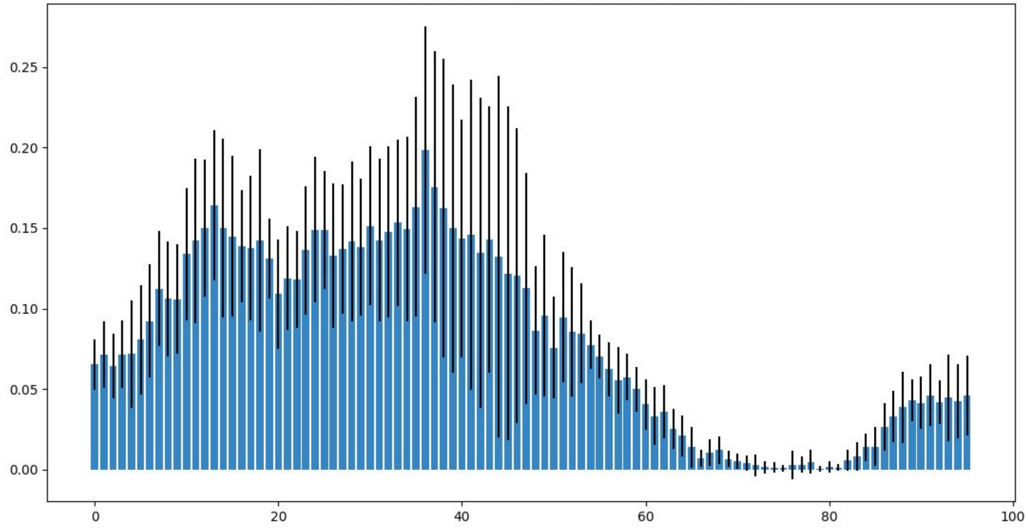
mean and sd of demand at: qp03rf during weekday: 4



mean and sd of demand at: qp03rf during weekday: 5

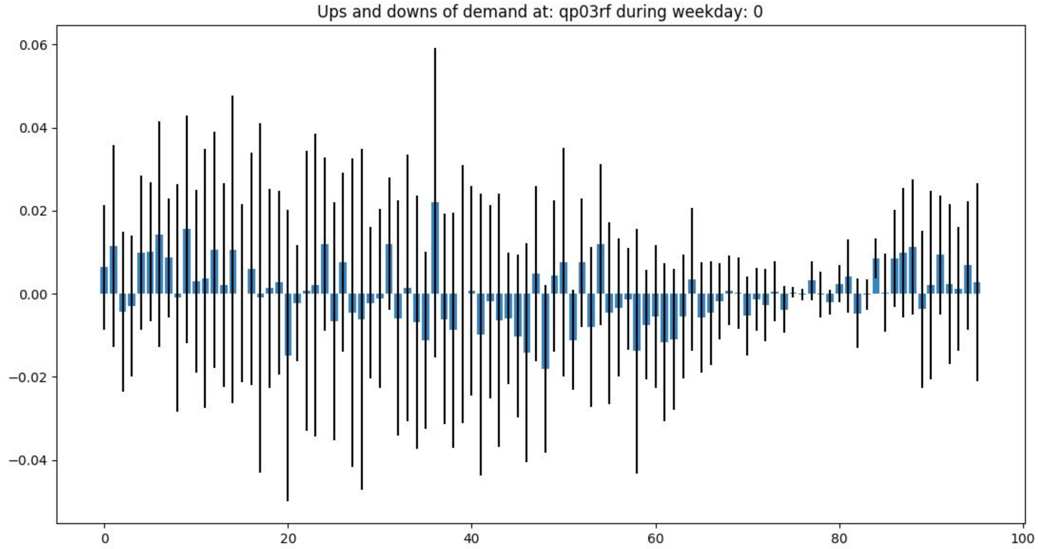


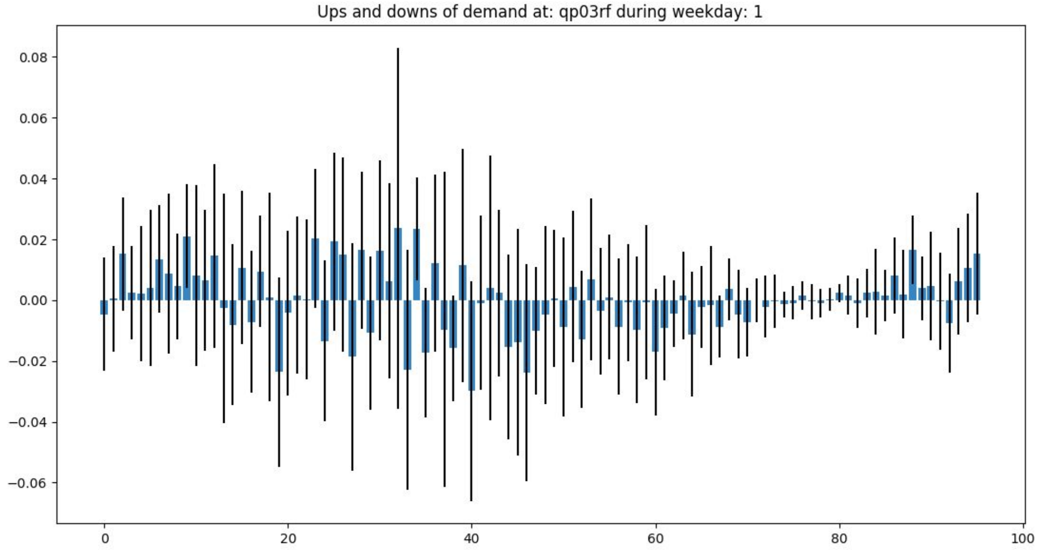
mean and sd of demand at: qp03rf during weekday: 6



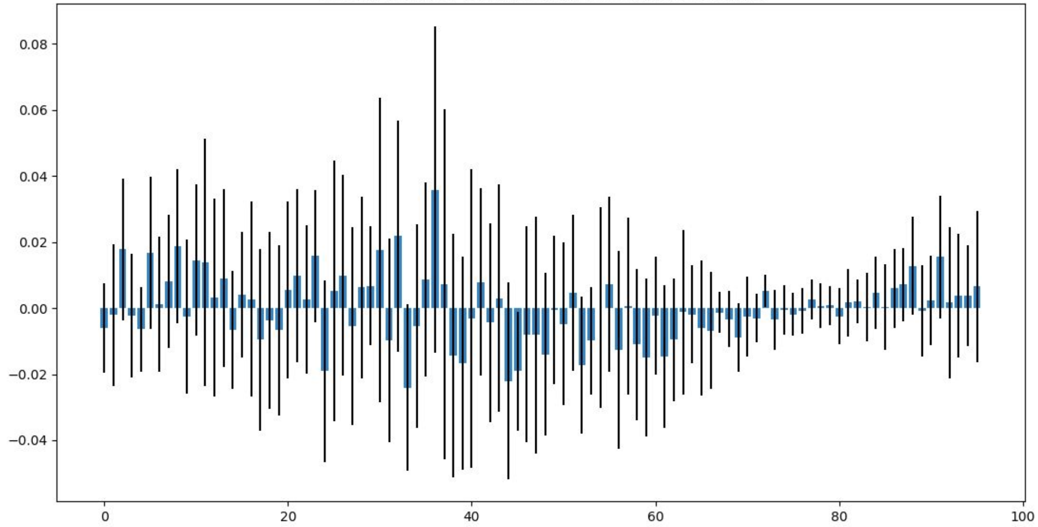
Although there is no significant difference between most of the figures above, it is fair to say that the first and the last two figures exhibit different patterns than the rest. Similar findings that are found through the repetition of this type of analysis using other random locations concludes that it is necessary to characterize the demand pattern not only by the location, but also by the weekday.

Is there any pattern in the change of demand from one moment to the next?
This question kinda makes sense as it is reasonable to imagine that the demand tends to increase at certain hours in a day and decrease at another hours. The figures below shows what is happening to the change in demand from time to time at a few locations for given weekday. The height of the bar represents the difference between the demand at one time and the demand at the next 15 minutes, while the black line represents the standard deviation of that difference.

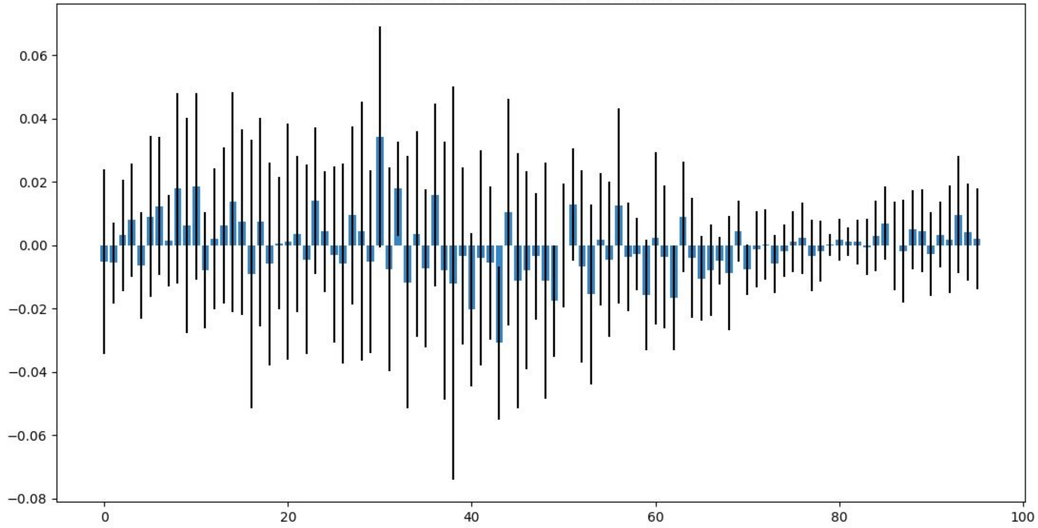




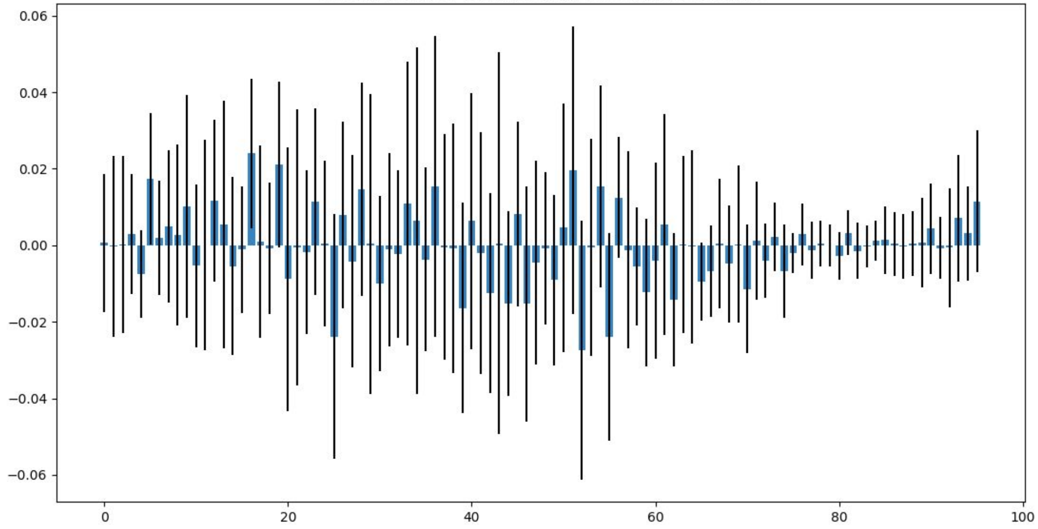
Ups and downs of demand at: qp03rf during weekday: 2

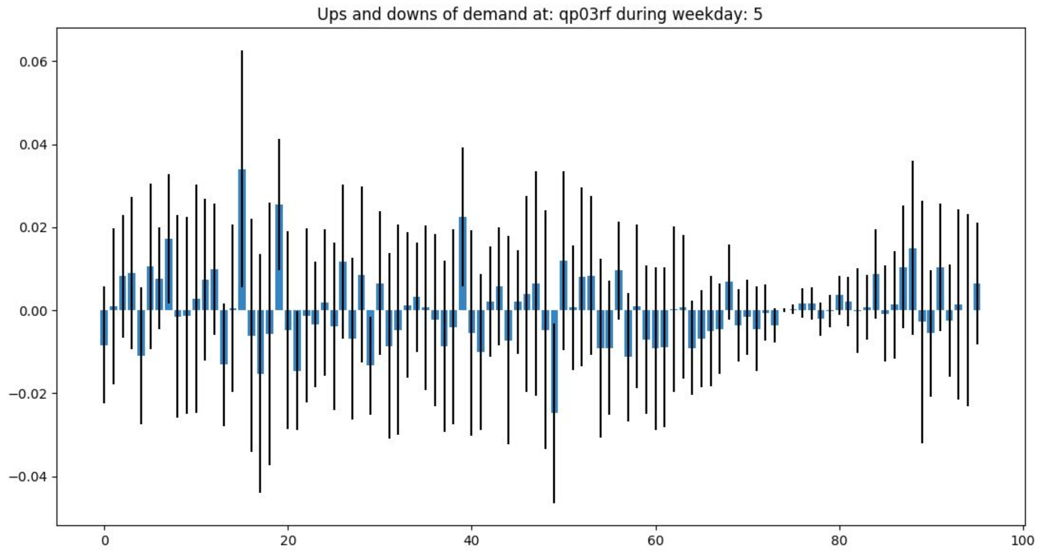


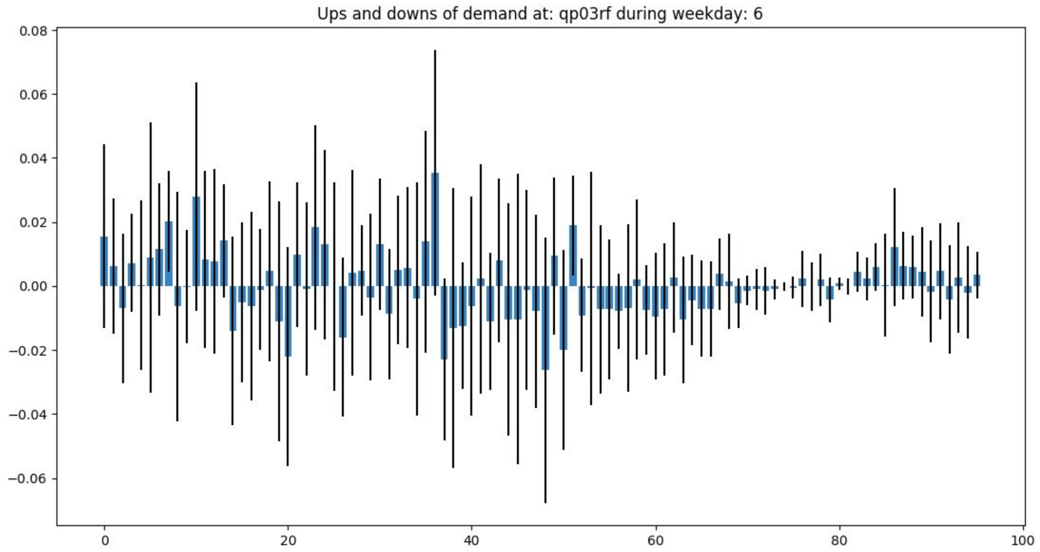
Ups and downs of demand at: qp03rf during weekday: 3



Ups and downs of demand at: qp03rf during weekday: 4



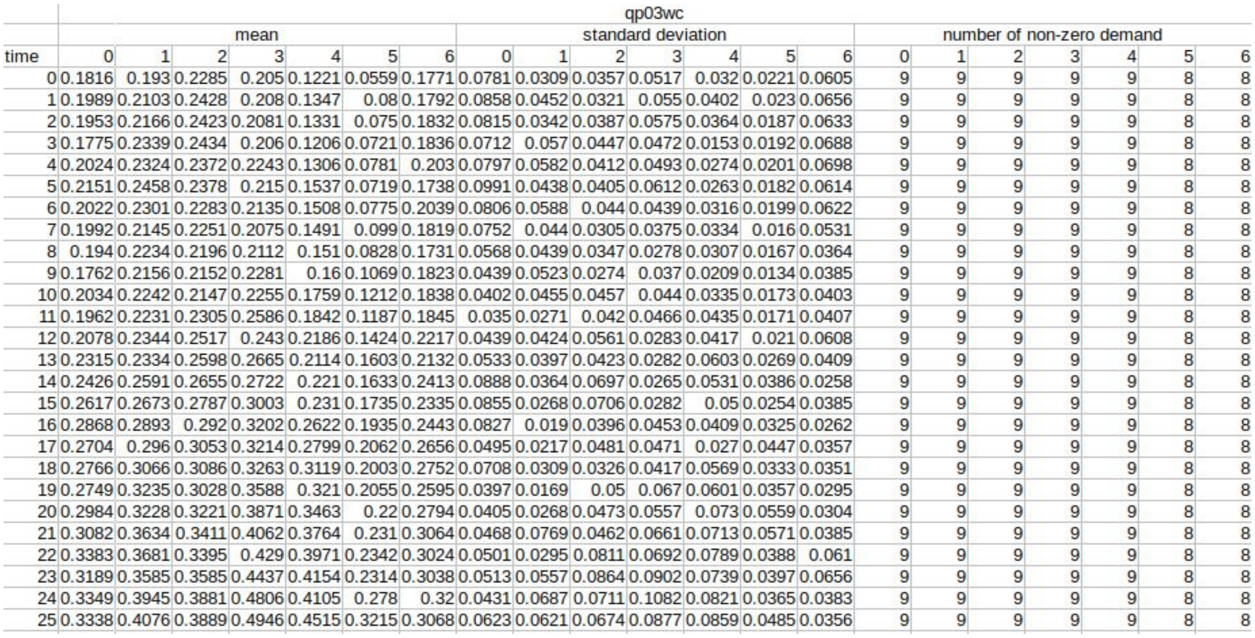




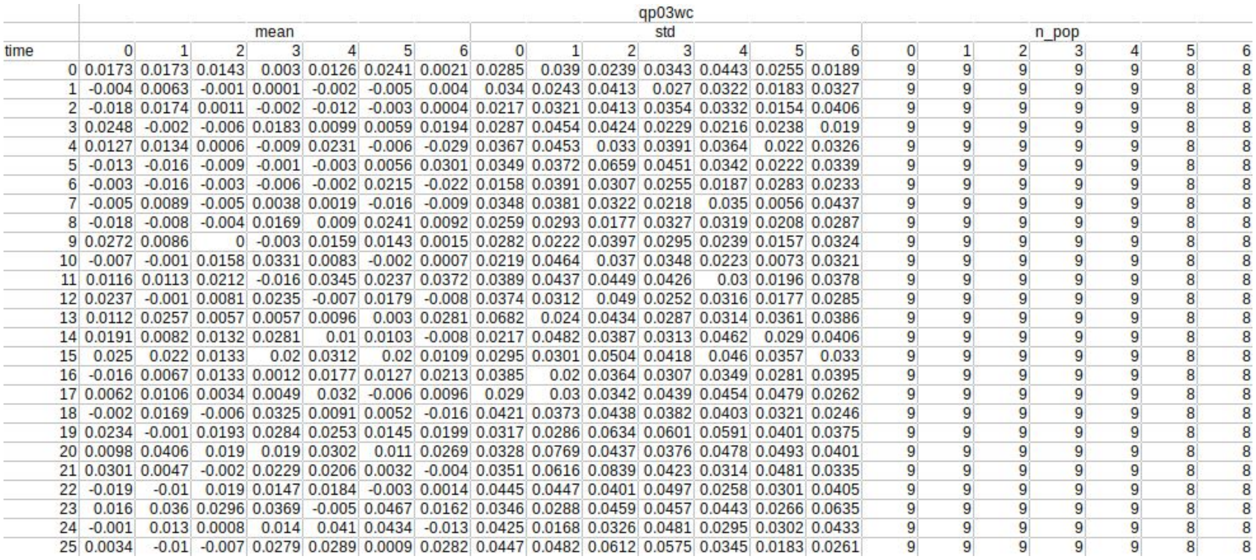
Due to the large standard deviation value, it is inconclusive to say whether there is a pattern in the change of demand from time to time. However, this information is still worth considering during the development of the forecasting model. Perhaps the machine can figure out what is going on with the demand change.

4. Constructing the location profile

Instead of using the demand profile (which classify any weekday as uniform), I devised a ‘location profile’. The location profile is the embodiment of the demand pattern at a certain location, weekday and time. Based on the analyses in the previous section, the location profile of each location consists of the mean and standard deviation of the demand at each unique time in a day and each weekday, and the mean and standard deviation of change in demand at each unique time in a day and each weekday, with the addition of the number of non-zero demand corresponding time and weekday. The advantages of using the location profile is that it is human friendly, it reduces a lot of complexity that the subsequent forecasting model has to deal with, it is modular as we can apply other types of forecasting model or update the location profile without requires major changes in the subsequent forecasting model, and most importantly, it is scalable as we can add/remove more locations. Below is one example of the location profile for location with geohash6 = ‘qp03wc’, where the numbers in the third row represent the weekday index (0 - 6). The rows shown in the figure are incomplete (in total there are 96 rows in total, representing unique time in a day). The first figure contains the statistical properties of the demand, and the second figure contains the statistical properties of the change in demand.



	qp03wc																				
	mean							standard deviation							number of non-zero demand						
time	0	1	2	3	4	5	6	0	1	2	3	4	5	6	0	1	2	3	4	5	6
0	0.1816	0.193	0.2285	0.205	0.1221	0.0559	0.1771	0.0781	0.0309	0.0357	0.0517	0.032	0.0221	0.0605	9	9	9	9	9	8	8
1	0.1989	0.2103	0.2428	0.208	0.1347	0.08	0.1792	0.0858	0.0452	0.0321	0.055	0.0402	0.023	0.0656	9	9	9	9	9	8	8
2	0.1953	0.2166	0.2423	0.2081	0.1331	0.075	0.1832	0.0815	0.0342	0.0387	0.0575	0.0364	0.0187	0.0633	9	9	9	9	9	8	8
3	0.1775	0.2339	0.2434	0.206	0.1206	0.0721	0.1836	0.0712	0.057	0.0447	0.0472	0.0153	0.0192	0.0688	9	9	9	9	9	8	8
4	0.2024	0.2324	0.2372	0.2243	0.1306	0.0781	0.203	0.0797	0.0582	0.0412	0.0493	0.0274	0.0201	0.0698	9	9	9	9	9	8	8
5	0.2151	0.2458	0.2378	0.215	0.1537	0.0719	0.1738	0.0991	0.0438	0.0405	0.0612	0.0263	0.0182	0.0614	9	9	9	9	9	8	8
6	0.2022	0.2301	0.2283	0.2135	0.1508	0.0775	0.2039	0.0806	0.0588	0.044	0.0439	0.0316	0.0199	0.0622	9	9	9	9	9	8	8
7	0.1992	0.2145	0.2251	0.2075	0.1491	0.099	0.1819	0.0752	0.044	0.0305	0.0375	0.0334	0.016	0.0531	9	9	9	9	9	8	8
8	0.194	0.2234	0.2196	0.2112	0.151	0.0828	0.1731	0.0568	0.0439	0.0347	0.0278	0.0307	0.0167	0.0364	9	9	9	9	9	8	8
9	0.1762	0.2156	0.2152	0.2281	0.16	0.1069	0.1823	0.0439	0.0523	0.0274	0.037	0.0209	0.0134	0.0385	9	9	9	9	9	8	8
10	0.2034	0.2242	0.2147	0.2255	0.1759	0.1212	0.1838	0.0402	0.0455	0.0457	0.044	0.0335	0.0173	0.0403	9	9	9	9	9	8	8
11	0.1962	0.2231	0.2305	0.2586	0.1842	0.1187	0.1845	0.035	0.0271	0.042	0.0466	0.0435	0.0171	0.0407	9	9	9	9	9	8	8
12	0.2078	0.2344	0.2517	0.243	0.2186	0.1424	0.2217	0.0439	0.0424	0.0561	0.0283	0.0417	0.021	0.0608	9	9	9	9	9	8	8
13	0.2315	0.2334	0.2598	0.2665	0.2114	0.1603	0.2132	0.0533	0.0397	0.0423	0.0282	0.0603	0.0269	0.0409	9	9	9	9	9	8	8
14	0.2426	0.2591	0.2655	0.2722	0.221	0.1633	0.2413	0.0888	0.0364	0.0697	0.0265	0.0531	0.0386	0.0258	9	9	9	9	9	8	8
15	0.2617	0.2673	0.2787	0.3003	0.231	0.1735	0.2335	0.0855	0.0268	0.0706	0.0282	0.05	0.0254	0.0385	9	9	9	9	9	8	8
16	0.2868	0.2893	0.292	0.3202	0.2622	0.1935	0.2443	0.0827	0.019	0.0396	0.0453	0.0409	0.0325	0.0262	9	9	9	9	9	8	8
17	0.2704	0.296	0.3053	0.3214	0.2799	0.2062	0.2656	0.0495	0.0217	0.0481	0.0471	0.027	0.0447	0.0357	9	9	9	9	9	8	8
18	0.2766	0.3066	0.3086	0.3263	0.3119	0.2003	0.2752	0.0708	0.0309	0.0326	0.0417	0.0569	0.0333	0.0351	9	9	9	9	9	8	8
19	0.2749	0.3235	0.3028	0.3588	0.321	0.2055	0.2595	0.0397	0.0169	0.05	0.067	0.0601	0.0357	0.0295	9	9	9	9	9	8	8
20	0.2984	0.3228	0.3221	0.3871	0.3463	0.22	0.2794	0.0405	0.0268	0.0473	0.0557	0.073	0.0559	0.0304	9	9	9	9	9	8	8
21	0.3082	0.3634	0.3411	0.4062	0.3764	0.231	0.3064	0.0468	0.0769	0.0462	0.0661	0.0713	0.0571	0.0385	9	9	9	9	9	8	8
22	0.3383	0.3681	0.3395	0.429	0.3971	0.2342	0.3024	0.0501	0.0295	0.0811	0.0692	0.0789	0.0388	0.061	9	9	9	9	9	8	8
23	0.3189	0.3585	0.3585	0.4437	0.4154	0.2314	0.3038	0.0513	0.0557	0.0864	0.0902	0.0739	0.0397	0.0656	9	9	9	9	9	8	8
24	0.3349	0.3945	0.3881	0.4806	0.4105	0.278	0.32	0.0431	0.0687	0.0711	0.1082	0.0821	0.0365	0.0383	9	9	9	9	9	8	8
25	0.3338	0.4076	0.3889	0.4946	0.4515	0.3215	0.3068	0.0623	0.0621	0.0674	0.0877	0.0859	0.0485	0.0356	9	9	9	9	9	8	8



	qp03wc																				
	mean							std							n_pop						
time	0	1	2	3	4	5	6	0	1	2	3	4	5	6	0	1	2	3	4	5	6
0	0.0173	0.0173	0.0143	0.003	0.0126	0.0241	0.0021	0.0285	0.039	0.0239	0.0343	0.0443	0.0255	0.0189	9	9	9	9	9	8	8
1	-0.004	0.0063	-0.001	0.0001	-0.002	-0.005	0.004	0.034	0.0243	0.0413	0.027	0.0322	0.0183	0.0327	9	9	9	9	9	8	8
2	-0.018	0.0174	0.0011	-0.002	-0.012	-0.003	0.0004	0.0217	0.0321	0.0413	0.0354	0.0332	0.0154	0.0406	9	9	9	9	9	8	8
3	0.0248	-0.002	-0.006	0.0183	0.0099	0.0059	0.0194	0.0287	0.0454	0.0424	0.0229	0.0216	0.0238	0.019	9	9	9	9	9	8	8
4	0.0127	0.0134	0.0006	-0.009	0.0231	-0.006	-0.029	0.0367	0.0453	0.033	0.0391	0.0364	0.022	0.0326	9	9	9	9	9	8	8
5	-0.013	-0.016	-0.009	-0.001	-0.003	0.0056	0.0301	0.0349	0.0372	0.0659	0.0451	0.0342	0.0222	0.0339	9	9	9	9	9	8	8
6	-0.003	-0.016	-0.003	-0.006	-0.002	0.0215	-0.022	0.0158	0.0391	0.0307	0.0255	0.0187	0.0283	0.0233	9	9	9	9	9	8	8
7	-0.005	0.0089	-0.005	0.0038	0.0019	-0.016	-0.009	0.0348	0.0381	0.0322	0.0218	0.035	0.0056	0.0437	9	9	9	9	9	8	8
8	-0.018	-0.008	-0.004	0.0169	0.009	0.0241	0.0092	0.0259	0.0293	0.0177	0.0327	0.0319	0.0208	0.0287	9	9	9	9	9	8	8
9	0.0272	0.0086	0	-0.003	0.0159	0.0143	0.0015	0.0282	0.0222	0.0397	0.0295	0.0239	0.0157	0.0324	9	9	9	9	9	8	8
10	-0.007	-0.001	0.0158	0.0331	0.0083	-0.002	0.0007	0.0219	0.0464	0.037	0.0348	0.0223	0.0073	0.0321	9	9	9	9	9	8	8
11	0.0116	0.0113	0.0212	-0.016	0.0345	0.0237	0.0372	0.0389	0.0437	0.0449	0.0426	0.03	0.0196	0.0378	9	9	9	9	9	8	8
12	0.0237	-0.001	0.0081	0.0235	-0.007	0.0179	-0.008	0.0374	0.0312	0.049	0.0252	0.0316	0.0177	0.0285	9	9	9	9	9	8	8
13	0.0112	0.0257	0.0057	0.0057	0.0096	0.003	0.0281	0.0682	0.024	0.0434	0.0287	0.0314	0.0361	0.0386	9	9	9	9	9	8	8
14	0.0191	0.0082	0.0132	0.0281	0.01	0.0103	-0.008	0.0217	0.0482	0.0387	0.0313	0.0462	0.029	0.0406	9	9	9	9	9	8	8
15	0.025	0.022	0.0133	0.02	0.0312	0.02	0.0109	0.0295	0.0301	0.0504	0.0418	0.046	0.0357	0.033	9	9	9	9	9	8	8
16	-0.016	0.0067	0.0133	0.0012	0.0177	0.0127	0.0213	0.0385	0.02	0.0364	0.0307	0.0349	0.0281	0.0395	9	9	9	9	9	8	8
17	0.0062	0.0106	0.0034	0.0049	0.032	-0.006	0.0096	0.029	0.03	0.0342	0.0439	0.0454	0.0479	0.0262	9	9	9	9	9	8	8
18	-0.002	0.0169	-0.006	0.0325	0.0091	0.0052	-0.016	0.0421	0.0373	0.0438	0.0382	0.0403	0.0321	0.0246	9	9	9	9	9	8	8
19	0.0234	-0.001	0.0193	0.0284	0.0253	0.0145	0.0199	0.0317	0.0286	0.0634	0.0601	0.0591	0.0401	0.0375	9	9	9	9	9	8	8
20	0.0098	0.0406	0.019	0.019	0.0302	0.011	0.0269	0.0328	0.0769	0.0437	0.0376	0.0478	0.0493	0.0401	9	9	9	9	9	8	8
21	0.0301	0.0047	-0.002	0.0229	0.0206	0.0032	-0.004	0.0351	0.0616	0.0839	0.0423	0.0314	0.0481	0.0335	9	9	9	9	9	8	8
22	-0.019	-0.01	0.019	0.0147	0.0184	-0.003	0.0014	0.0445	0.0447	0.0401	0.0497	0.0258	0.0301	0.0405	9	9	9	9	9	8	8
23	0.016	0.036	0.0296	0.0369	-0.005	0.0467	0.0162	0.0346	0.0288	0.0459	0.0457	0.0443	0.0266	0.0635	9	9	9	9	9	8	8
24	-0.001	0.013	0.0008	0.014	0.041	0.0434	-0.013	0.0425	0.0168	0.0326	0.0481	0.0295	0.0302	0.0433	9	9	9	9	9	8	8
25	0.0034	-0.01	-0.007	0.0279	0.0289	0.0009	0.0282	0.0447	0.0482	0.0612	0.0575	0.0345	0.0183	0.0261	9	9	9	9	9	8	8

This location profile alone can already acts as the forecasting model with reasonable accuracy. However, more can be done using new technology such as artificial neural network.

5. Designing the forecasting model

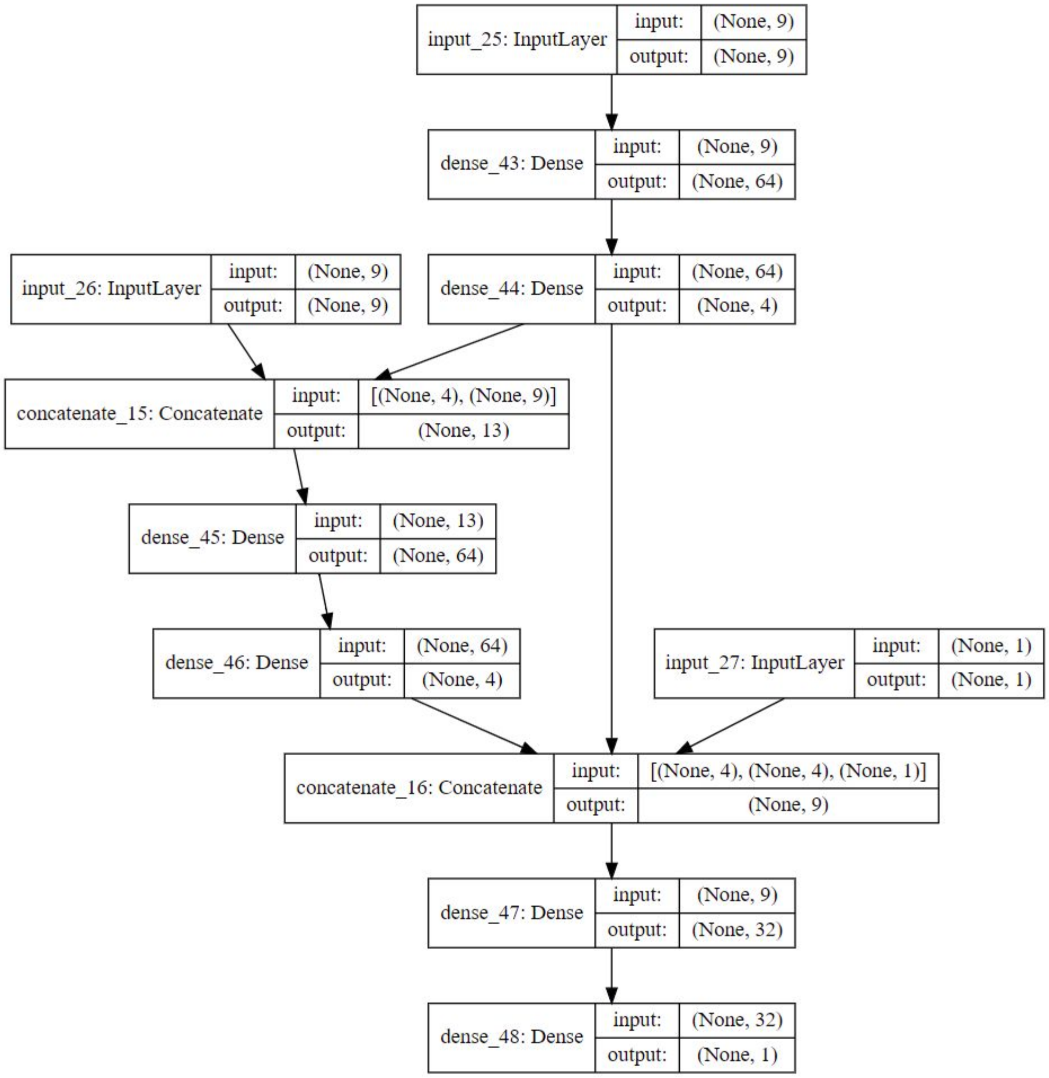
After comparing the convolutional neural network (CNN), simple feed forward neural network (NN) and lightBGM. I decided to use NN because it offers easier deployment and maintenance, and the accuracy of the NN is on par with the lightBGM. Meanwhile, the CNN requires too much computing power and the resulting accuracy after limited time of training is still not as good as that of the lightBGM. The obvious question when designing the NN is what are the features? And what is the architecture of the NN.

What are the features?

There are two important information that we have discovered so far, the pattern of demands as represented by the location profile, and the demand itself. Therefore, the features must include both information.

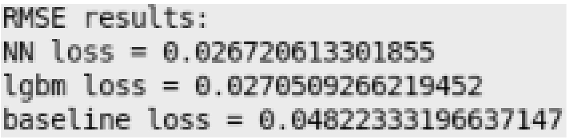
What is the architecture of the NN?

The neural network contains three input layers and consists of reasonably small amount of neurons. Below is a figure showing the architecture of the NN. The first input layer (input_25) receives 9 features data obtained from the location profile. These data are: mean of the demand at user inputted timestamp +0 minutes, +15 minutes and +30 minutes; standard deviation of the demand at user inputted timestamp +0 minutes, +15 minutes and +30 minutes; and normalized number of non-zero demand at user inputted timestamp +0 minutes, +15 minutes and + 30 minutes. The second input layer (input_26) also receives 9 features data obtained from the location profile. These data are: mean of demand at user inputted timestamp +0 minutes, +15 minutes and +30 minutes; standard deviation of the change in demand at user inputted timestamp +0 minutes, +15 minutes and +30 minutes; and normalized number of non-zero demand at user inputted timestamp +0 minutes, +15 minutes and + 30 minutes. Finally, the last input layer, receives the value of user inputted demand. Then, the neural network will return the forecasted demand at the user inputted location during the next 15 minutes.
The intuition behind the proposed NN architecture is that we want the NN to understand the rough overview of how the demand looks like in the past. Then, given the historical demand pattern information, we want the NN to also consider the uncertainty in demand changes over time. Finally, give all those knowledge, we give the NN the information about current demand and let the NN draws the conclusion considering the past and present information.

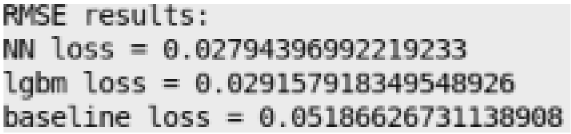


6. Validation of the forecasting model

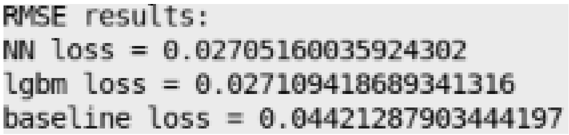
The validation of the proposed forecasting model (NN) is done by comparing the NN model with the baseline model (based purely on the location profile; baseline) and the lightbgm-based (lgbm) model. Below are the five random validation results for T + 1 prediction, each using 1000 random training data and measured by the root mean squared error (RMSE).



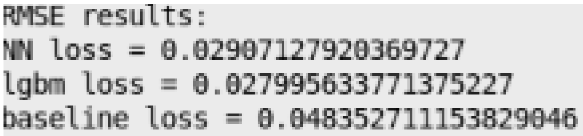
```
RMSE results:
NN loss = 0.026720613301855
lgbm loss = 0.0270509266219452
baseline loss = 0.04822333196637147
```



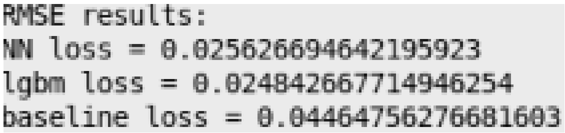
```
RMSE results:
NN loss = 0.02794396992219233
lgbm loss = 0.029157918349548926
baseline loss = 0.05186626731138908
```



```
RMSE results:
NN loss = 0.02705160035924302
lgbm loss = 0.027109418689341316
baseline loss = 0.04421287903444197
```



```
RMSE results:
NN loss = 0.02907127920369727
lgbm loss = 0.027995633771375227
baseline loss = 0.048352711153829046
```



```
RMSE results:
NN loss = 0.025626694642195923
lgbm loss = 0.024842667714946254
baseline loss = 0.04464756276681603
```

Notice that although the NN still lags behind the lbgm at some cases, in the long run, NN is more suitable because of the easy continuous re-training based on data update that must be conducted to adapt the changing demand.

7. Deployment and sustainability of the forecasting model

Once the forecasting model has been trained and saved, it is ready to be used in day to day operation. The model can be used to forecast the demand during the next 15 minutes at certain location given the current demand and the location profile. The more challenging issue is how the model can adapt to the changing demand over time.

How the model can adapt to the changing demand over time?
Although adapting the model itself is crucial, what is just as important is to ensure that the features used during the adaptation process is still relevant. Since the information used to generate the features are stored in the location profile, we have to first update the location profile before updating the model. I suggest to perform the update on the location profile every week (per 7 days) using the demand data within the timespan of the last 4 to 8 weeks. The decision on the timespan depends on whether there was any special event that significantly change the demand pattern and must be ignored, or whether there was nothing much happen during the last few weeks so that we have more demand data to be considered during the profile update which results in a more robust location profile. For instance, in this week, we can use the demand data for the last 8 weeks if there was nothing much happen, no Chinese New Year, no Christmas, just casual days. However, if in the last two weeks the demand were severely affected by random major event. We may use the last 8 weeks data minus the recent 2 weeks data, which in total becomes a 6 weeks data to update the location profile. Finally, after we have the update location profile, we can adapt the NN model to the changing

demand through re-training using the updated location profile and demand data used during location profile update.

8. Conclusion and remarks

In this report, we have explored the patterns that emerges in the demand at various locations and time, developed a location profile to store the characteristics of the demand patterns, built a forecasting model, and finally discussed how the model can be deployed and maintained to adapt to the ever changing world. Although the concept seems sensible and easy to implement, there are still more research that must be conducted to improve the presented concept. For instance, in the long run, we can maintain more than a single location profile to account for unique yet persistent and predictable demand change, such as during special event in which the demand is expected to significantly increases/decreases. Another improvement can also be done by adding a mechanism to account for user behavioural change once the concept is implemented in such a way that the behaviour of the user (transportation service provider) changes as they have access to the forecasted demand information. In this scenario, the forecasting model may tend to be ‘wrong’ most of the time due to bias introduced by the expected demand. For instance, users flock to where the demand is expected to increase, therefore the demand at the less covered area accumulate more than usual. Meanwhile, when the user has no information of the future expected demand, the forecasting model works just fine. Another thing to consider is to include more features to predict the demand such as the real weather and the weather forecast information which I believe affects people's decision in how to commute. However, based on the validation results, the concept presented above is still remains a great choice at the moment to forecast future demand.