Introduction to Data Science course, fall 2020

Technical report of the mini-project

Helsinki cycling data and weather

Kari Ojala and Tapio Koukkari

20.10.2020

Original pitch of the project from Sep 13, 2020

Helsinki is investing a lot in new bicycle routes and city bikes. How weather affects cycling is not that much discussed. Combining cycling data with weather data should give interesting results for e.g. bicycle shops, city planners and to cyclists themselves. There is a huge amount of data available. Challenges are to find the most relevant data combinations and their effective visualization.

Tidying data

Originally, we had a third member in the group: Laura Aarnio. These datasets were agreed with her.

- Weather data in Helsinki Kaisaniemi in 2019, from https://www.ilmatieteenlaitos.fi/avoin-data-avattavat-aineistot
- Cycling data in Helsinki Kaisaniemi, all bikes in 2019, from https://www.avoindata.fi/data/fi/dataset/helsingin-pyorailijamaarat
- City bike data in Helsinki Kaisaniemi city bicycle station 3.4.-31.10.2019, from https://www.avoindata.fi/data/fi/dataset/helsingin-ja-espoon-kaupunkipyorilla-ajatut-matkat

Hourly data from these datasets were combined into two datasets. File **bikeweather-2019-insummertime.csv** contains weather data and cycling data of all bikes from the whole year. File **kaisaniemi_summer_combined.csv** contains hourly data from all three sources for the period when city bikes were available.

The following tidying actions were done on the two datasets:

- The weather data was changed from UTC time to Finland summer time, to be compatible with bike data.
- In weather data, values of -1 in rain data (mm/hour) and snow depth data (cm) were changed to 0.
- In original weather data, missing cloudiness data was marked with value 9. These were changed first into NaN-values, which were then forward filled using pandas method ‘ffill’.
- Citybike data (bikes available, spaces available) was collected in 10 min intervals. These values were changed into hourly averages.

Tidying was carried out by using varying python codes and methods (not included in this report).

Data Analysis tools

Data analysis files are available in https://github.com/ktojala/DataScience_project_2020. Jupyter Notebook reports for the basic data analysis **(bikeweatherstudy.ipynb)** and widget **(widget.ipynb)** are available in the GIT-page. Relatively simple pandas methods were used for the analysis of bike data and weather data. Analysis could be upgraded by using for example the plotly library from https://plotly.com/python/getting-started/.

General biking analysis

This preliminary study targets to identify the key factors and features that should be taken into account, when predicting biking in general and after that, the use of city bikes. It is important to note that all data in this report is Kaisaniemi specific and does not necessarily apply as such for other bike stations. Location of the station itself is one of the key factors, as can be seen in Figure 1.

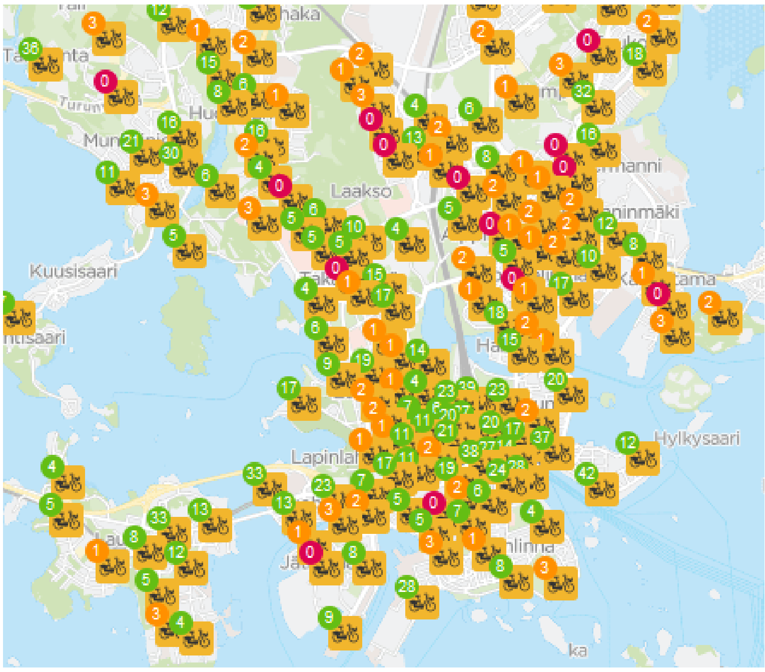


Figure 1. City bike data in Helsinki on October 20, 2019, at 11:35. Some bike stations are empty (red zeros), while most other stations have a lot of city bikes available (green numbers). Source: https://kaupunkipyorat.hsl.fi/fi/helsinki/stations.

Weather dependence, time dependence, and group dependence

Group dependencies are linked to time dependence. Different groups of people are biking at different times of the week and different times of the day.

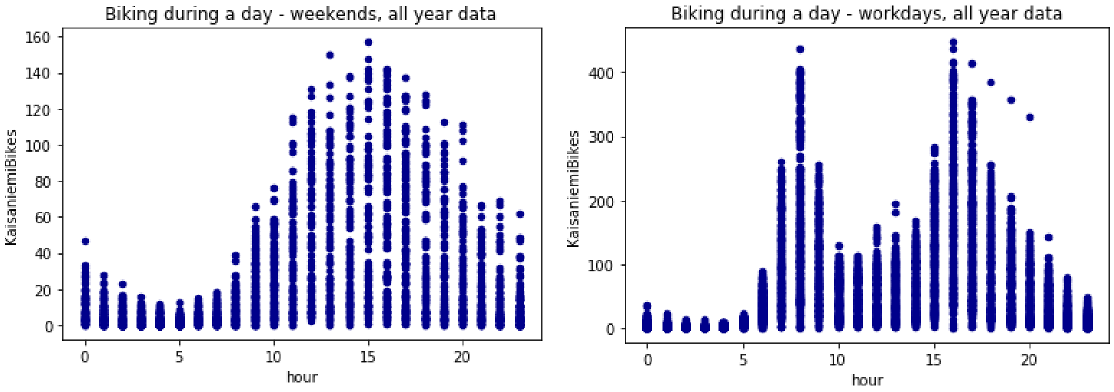


Figure 2. Weekend (left) bike traffic during daytime is very different from weekdays (right). Each dot in the plots represents the number of bikers during one hour period in Kaisaniemi.

We can see from Figure 2 that during workdays, there is relatively high bike traffic in the morning between 6-9 a.m. and 2-5 p.m. These groups may represent workers or students, who do not need to go to their workplaces or schools during weekends. During weekends, a very different traffic pattern emerges with maximum at 2-3 p.m. and minimum at 4-5 a.m. However, during the year, there are also days with practically no traffic for every hour. (25th of December is particularly quiet.)

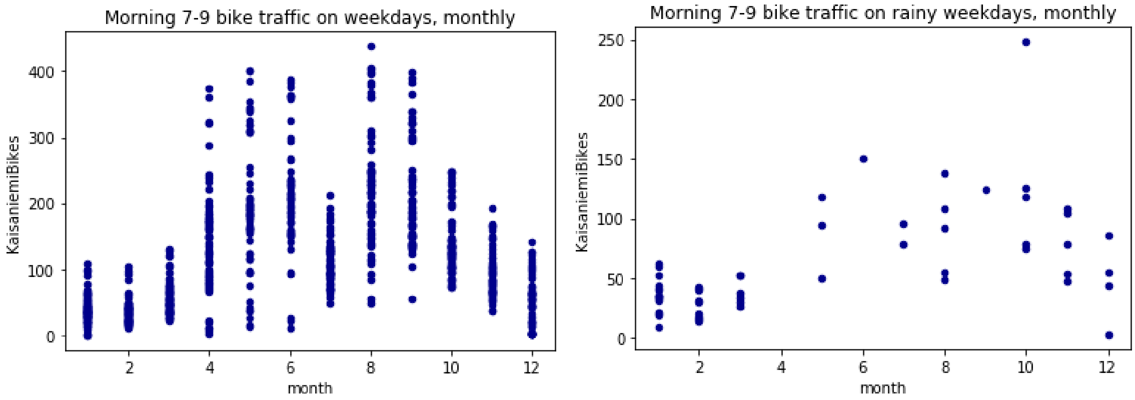


Figure 3. Weekday morning traffic between 6-9 a.m. during the year. In the left, all days are included, in the right, only rainy hours are. City bikes add to the traffic from April to October.

In Figure 3, we take a closer look at the morning traffic during the year. Holiday period in July is clearly visible with less morning traffic. As expected, winter period is quieter, but not without bike traffic. In the right, we can see that traffic during rainy hours is relatively quiet, but even then, there are bikers. Notice that there are quite few rainy mornings compared to the total number of hours.

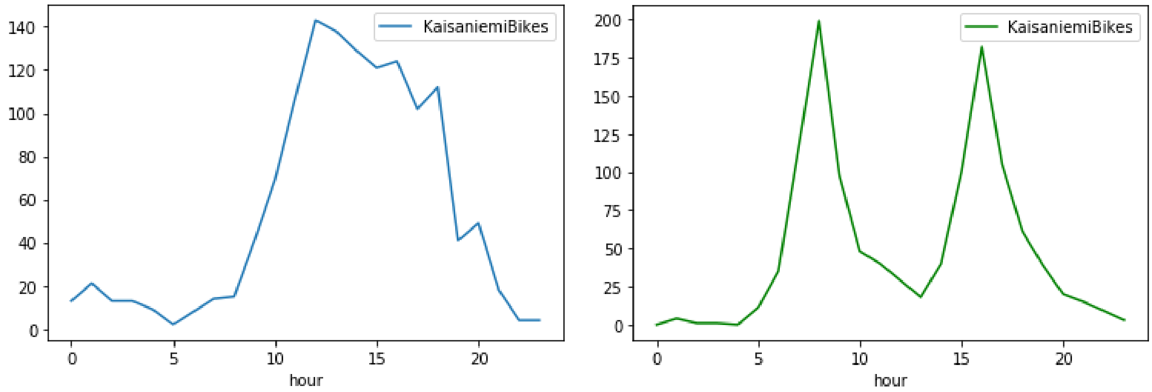


Figure 4. May 1st traffic (Wednesday) during the day on the left and May 2nd traffic on the right.

There are many particular days during the year, when traffic is not according to typical patterns. The first of May can be a weekday, but its bike traffic curve (Figure 4.) is similar to weekend traffic.

Which weather features are the most significant, when it comes to biking? A simple comparison using linear regression for each weather feature was carried out. Ar first, each feature at a time was selected as the explanatory variable, while “KaisaniemiBikes” was the response variable. After that, multivariable regression was done, adding one variable at a time into the list of feature variables. Results are collected into Table 1. It is clear that some features like wind direction are not suitable for linear regression. From these results we can infer, that the two most significant features are 1) temperature (ToC) and 2) relative humidity (RHpercent). Dewpoint of air (DewToC) looks signigicant, but is has a very high positive correlation 0,79 with air temperature ToC, so it was not selected. Temperature and relative humidity are an excellent pair, as their correlation is only 0,15. An interesting thing about relative humidity is that during the summer its correlation on biking was positive, while in winter, its correlation is negative. This can be explained so that in winter, dry air is usually related to cold temperature. In summmer, very humid air is related to rainy weather.

Table 1. (Multi)linear regression test for weather features on all hours of the year 2019.

lin. corr	Feature	R2 alone	R2 adds	Difference in R2	nonlinear !
neg	Clouds	0,04431	0,04431	0,04431	
pos	AirPhPa	0,02861	0,05513	0,01082	here
neg	Rhpercent	0,12768	0,12993	0,07480	
neg	RainmmH	0,00110	0,12998	0,00005	
neg	Snowcm	0,06159	0,15943	0,02945	
pos	ToC	0,20119	0,24120	0,08177	here
pos	DewToC	0,09598	0,24272	0,00152	
pos	Visibilitym	0,01405	0,24623	0,00351	
neg	Winddirdeg	0,00344	0,24710	0,00087	here (3 maxima)
pos	Gustms	0,00059	0,24801	0,00091	here
pos	Windms	0,00188	0,25093	0,00292	here

As Table 1 shows, all 11 weather features explain 25,1% of the number of bikers. The two most significant features alone explain the number of bikers almost as well: 24,1%. However, if we only look at weekends, those two features then explained 52,4% of the number of bikers in Kaisaniemi!

Some weather features have a clear effect on biking, but the dependence is non-linear. Temperature is one such feature, wind speed is another one, see Figure 5.

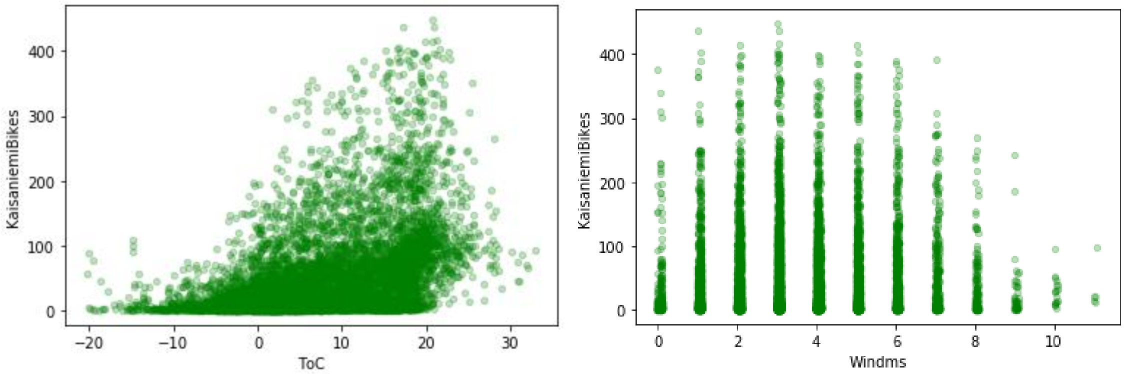
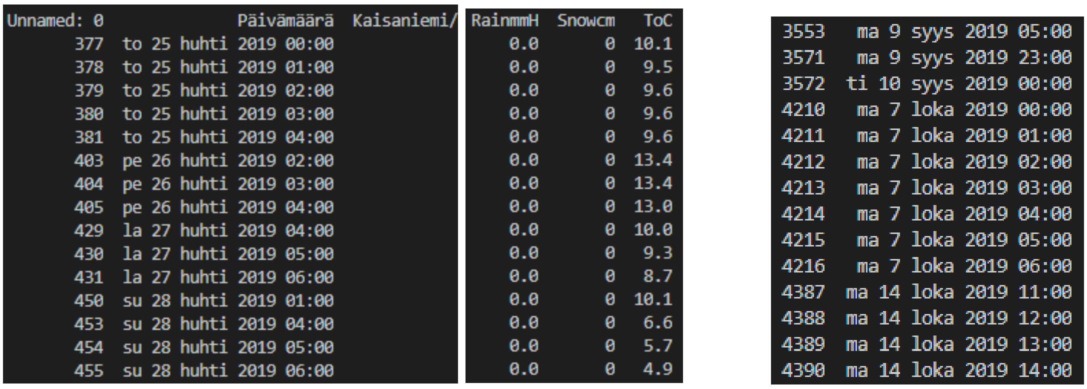


Figure 5. In some weather features, mimimum and maximum values are not preferred by bikers.

City bike data



```
Unnamed: 0             Päivämäärä  Kaisaniemi/
       377  to 25 huhti 2019 00:00
       378  to 25 huhti 2019 01:00
       379  to 25 huhti 2019 02:00
       380  to 25 huhti 2019 03:00
       381  to 25 huhti 2019 04:00
       403  pe 26 huhti 2019 02:00
       404  pe 26 huhti 2019 03:00
       405  pe 26 huhti 2019 04:00
       429  la 27 huhti 2019 04:00
       430  la 27 huhti 2019 05:00
       431  la 27 huhti 2019 06:00
       450  su 28 huhti 2019 01:00
       453  su 28 huhti 2019 04:00
       454  su 28 huhti 2019 05:00
       455  su 28 huhti 2019 06:00
```

```
RainmmH  Snowcm   ToC
    0.0       0  10.1
    0.0       0   9.5
    0.0       0   9.6
    0.0       0   9.6
    0.0       0   9.6
    0.0       0  13.4
    0.0       0  13.4
    0.0       0  13.0
    0.0       0  10.0
    0.0       0   9.3
    0.0       0   8.7
    0.0       0  10.1
    0.0       0   6.6
    0.0       0   5.7
    0.0       0   4.9
```

```
3553   ma 9 syys 2019 05:00
3571   ma 9 syys 2019 23:00
3572  ti 10 syys 2019 00:00
4210   ma 7 loka 2019 00:00
4211   ma 7 loka 2019 01:00
4212   ma 7 loka 2019 02:00
4213   ma 7 loka 2019 03:00
4214   ma 7 loka 2019 04:00
4215   ma 7 loka 2019 05:00
4216   ma 7 loka 2019 06:00
4387  ma 14 loka 2019 11:00
4388  ma 14 loka 2019 12:00
4389  ma 14 loka 2019 13:00
4390  ma 14 loka 2019 14:00
```

Figure 6. On the left, a typical time when all city bikes are in use in Kaisaniemi: weekend midnight, relatively warm weather and no rain. On the right, October 14 afternoon is an exception to the rule.

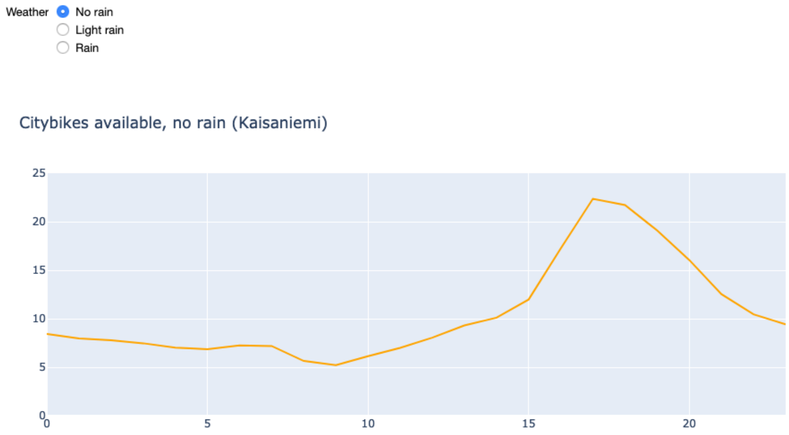
As can be seen in Figure 1, city bike usage depends on location. When exactly are there no city bikes available in Kaisaniemi? An interesting find was that weekend nights are especially critical for bike availability (Figure 6). A nice weather was an additional requirement. A probable explanation for this is that people come out from night clubs and there are no buses, trains or subway available.

There was one exception to the rule: Monday 14th of October, the first day of fall holiday week in Helsinki schools. In that afternoon, all city bikes from Kaisaniemi bike station were in use.

City bike predictive widget

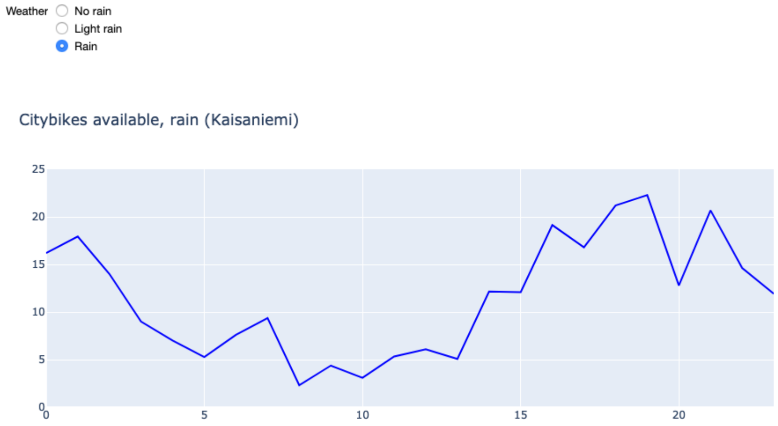
We also started to develop the bike availability prediction widget. Some preliminary results are shown in Figures 7-8. As there are really many factors affecting bike availability, it would require a rather extensive model to include all dependencies and codepencencies. We envision a tool, which would take the latest weather prediction as input, including predicted weather radar data. The tool would also take into account the fact that the number of bikers is probably increasing every year.

One day before the deadline of this report, it was found that Helsinki has just added a feature into its website https://www.findbikenow.com/#/map that predicts the availability of city bikes 1 hour or 24 hours ahead. It does not explain if weather is a factor in that prediction. Probably it is not.



Weather ● No rain
○ Light rain
○ Rain

Figure 7. Bikes available, no rain



Weather ○ No rain
○ Light rain
● Rain

Figure 8. Bikes available, rainy day

The widget, the Jupyter Notebook, the tidied datasets, this report and our pitch presentation on October 14, 2020 are available from https://github.com/ktojala/DataScience_project_2020.