Engeto_Data_Academy_Python_Project

January 28, 2022

1 Data Analysis of Bike Sharing in Edinburgh

Petr Vlasak

email: petr.vlasakk@gmail.com

gitHub: https://github.com/pvlasak

2 Input Dataframe Description

2.1 Dataframe Overview

Each line represents bike rental of the bike sharing service in Edinburgh.

```
   Unnamed: 0  index           started_at             ended_at  duration  \
0           0      0  2018-09-15 08:52:05  2018-09-15 09:11:48      1182
1           1      1  2018-09-15 09:24:33  2018-09-15 09:41:09       995
2           2      2  2018-09-15 09:48:54  2018-09-15 10:46:40      3466
3           3      3  2018-09-16 12:01:36  2018-09-16 12:25:26      1430
4           4      4  2018-09-16 12:03:43  2018-09-16 12:11:16       452

   start_station_id start_station_name         start_station_description  \
0               247   Charlotte Square  North Corner of Charlotte Square
1               259   St Andrew Square                 North East corner
2               262         Canonmills                      near Tesco's
3               255  Kings Buildings 4                          X-Y Cafe
4               255  Kings Buildings 4                          X-Y Cafe

   start_station_latitude  start_station_longitude  end_station_id  \
0               55.952335                -3.207101             259
1               55.954749                -3.192774             262
2               55.962804                -3.196284             250
3               55.922001                -3.176902             254
4               55.922001                -3.176902             253

   end_station_name                 end_station_description  \
0  St Andrew Square                       North East corner
1        Canonmills                            near Tesco's
2     Victoria Quay  Entrance to Scottish Government Office
```

```
3  Kings Building 3                    Kings Building House
4  Kings Building 2                      Sanderson Building

   end_station_latitude  end_station_longitude
0             55.954728             -3.192653
1             55.962804             -3.196284
2             55.977638             -3.174116
3             55.923479             -3.175385
4             55.923202             -3.171646
```

2.2 Basic statistical values of the dataframe

```
       Unnamed: 0      index    duration  start_station_id  \
count   438259.00  438259.00   438259.00         438259.00
mean    219129.00    9043.26     1948.84            924.25
std     126514.62    7439.18     5657.13            670.16
min          0.00       0.00       61.00            171.00
25%     109564.50    3252.00      624.00            260.00
50%     219129.00    7127.00     1163.00           1019.00
75%     328693.50   12467.00     2529.00           1728.00
max     438258.00   31397.00  2363348.00           2268.00

       start_station_latitude  start_station_longitude  end_station_id  \
count               438259.00                438259.00       438259.00
mean                    55.95                    -3.20          969.35
std                      0.01                     0.04          676.76
min                     55.91                    -3.41          171.00
25%                     55.94                    -3.21          262.00
50%                     55.95                    -3.19         1024.00
75%                     55.96                    -3.18         1737.00
max                     55.99                    -3.06         2268.00

       end_station_latitude  end_station_longitude
count             438259.00              438259.00
mean                  55.95                  -3.20
std                    0.02                   0.04
min                   53.40                  -3.41
25%                   55.94                  -3.21
50%                   55.95                  -3.19
75%                   55.96                  -3.18
max                   55.99                  -2.99
```

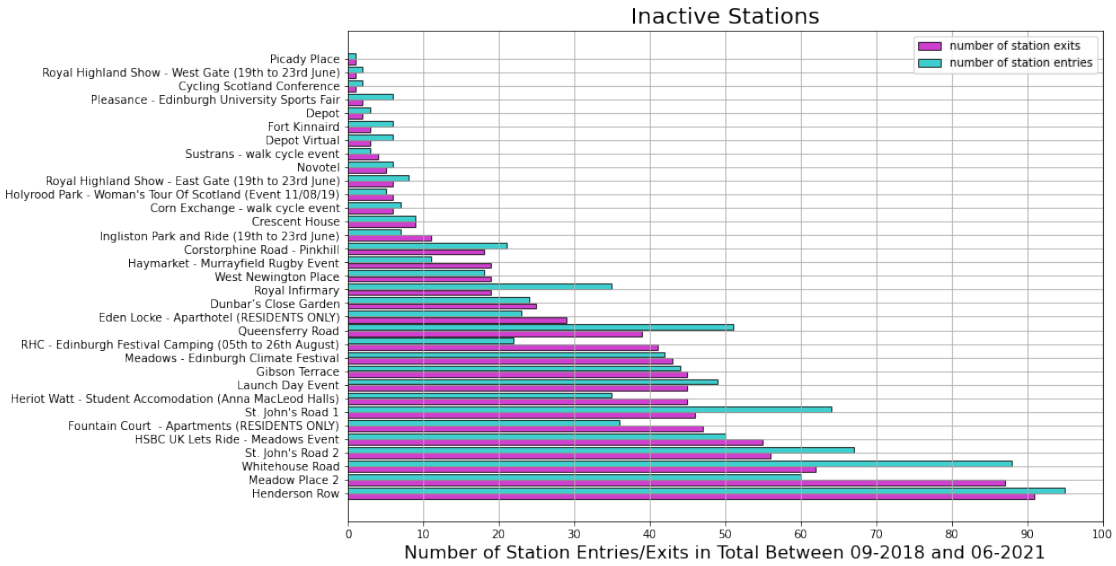
2.3 Dataframe Summary

a. Dataframe has 438259 entries in total.
b. Number of unique start stations is 169.
c. Number of unique end stations is 170.
d. First Bike Ride in the dataset started on 2018-09-15 08:52:05.

```
e. Last Bike Ride in the dataset started on 2021-06-30 23:58:33.
f. Bike ride takes in average 32.5 minutes.
g. The longest bike rental took 656.5 hours.
h. The shortest bike rental took 61 seconds.
i. Number of bike rentals with duration longer than 24 hours is :1.
j. Number of lines with NaN values per column:
----------------------------------------------------------------------
Unnamed: 0                       0
index                            0
started_at                       0
ended_at                         0
duration                         0
start_station_id                 0
start_station_name               0
start_station_description     4141
start_station_latitude           0
start_station_longitude          0
end_station_id                   0
end_station_name                 0
end_station_description       4689
end_station_latitude             0
end_station_longitude            0
dtype: int64
```

NaN values are located only in the column "station_description", which is less important information and therefore we can ignore these NaN values.

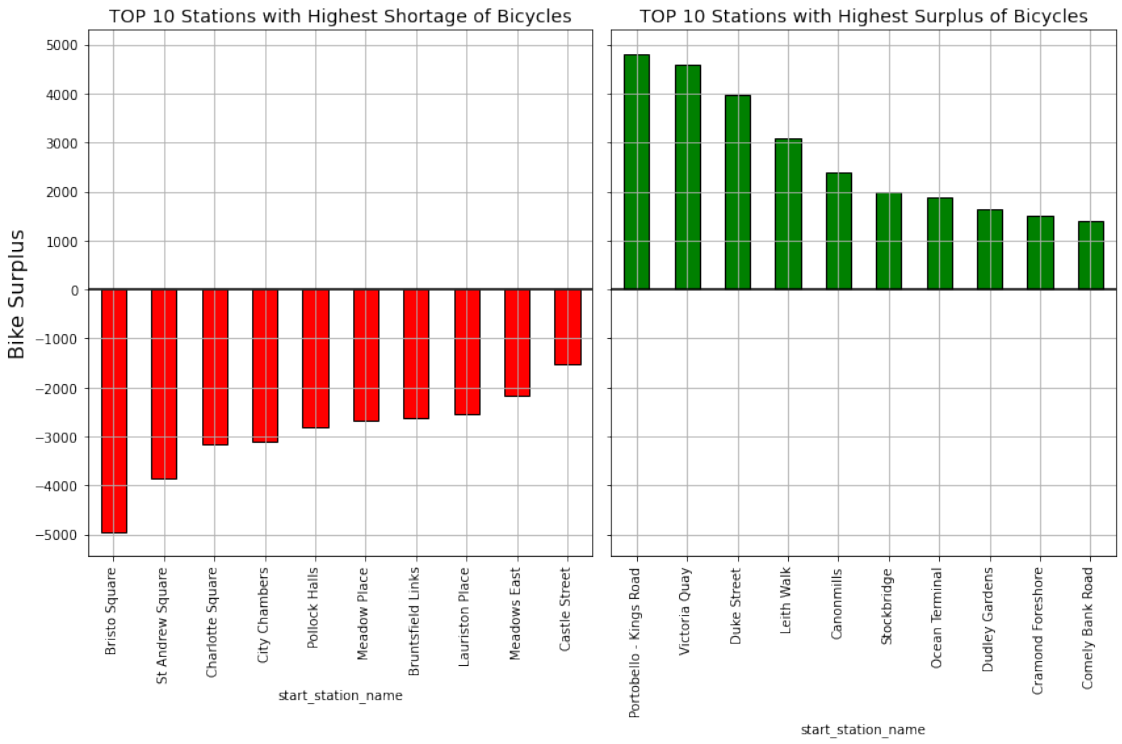
3 Inactive stations



Total number of station entries and exits has been calculated. Stations with both parameters lower than 100 were selected and plotted as a horizontal bar chart.

Conclusion : Picady Place, Royal Highland Show, Cycling Scotland Conference, Edinburgh University Sports Fair, Depot are 5 stations with the lowest number of visits and can be considered as inactive.

4 Bike Surplus and Shortage

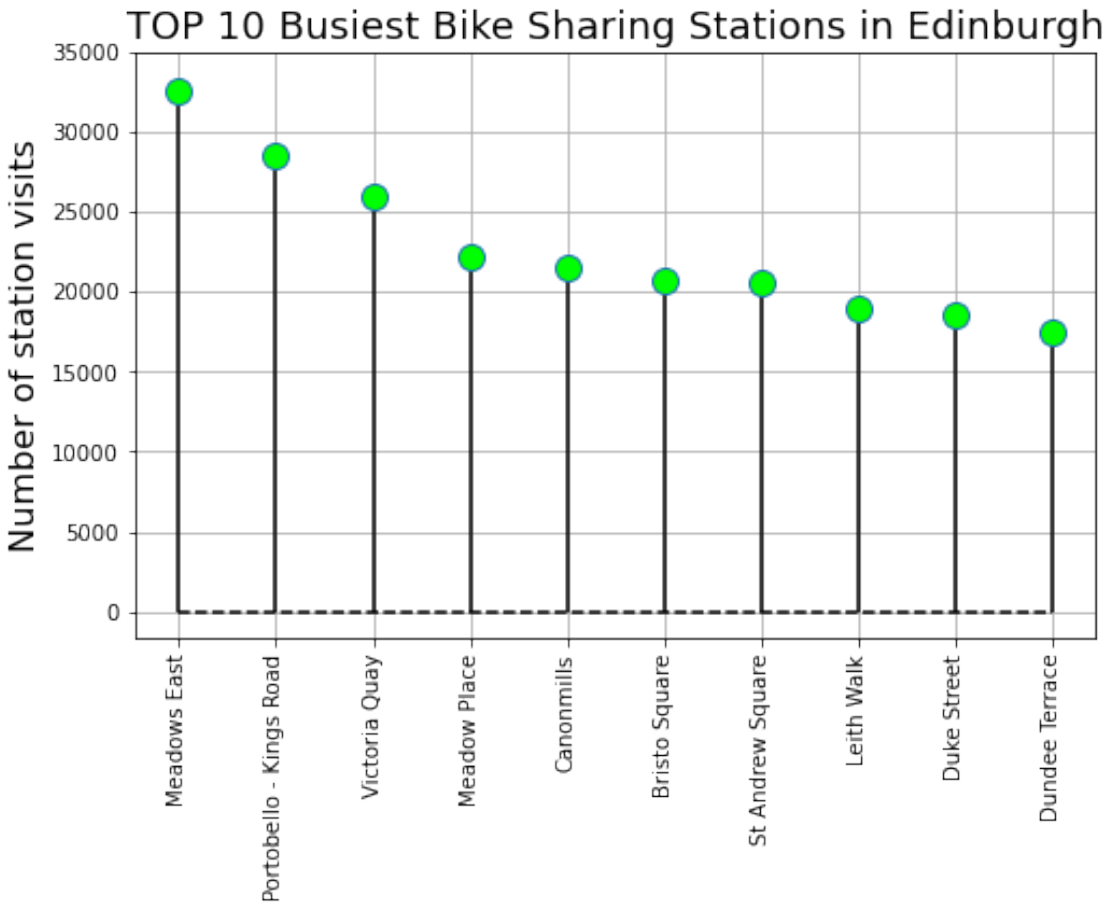


The goal was to find stations which in general suffer from a shortage of bikes or or vice versa from a surplus of bicycles.

Easiest way how to determine a bike shortage was to substract the number of exits from number of entries. Positive number indicates that the station has in general a surplus of bicycles, however negative value says that a bicycle shortage is typical for particular station.

Distadvange of this method is that it considers long-term data and daily or weekly bike shortage can´t be therefore identified. Daily or weekly bike shortage might be here more interesting output.

5 Busiest Stations



Busiest stations are identified by adding the sum of exits and sum of entries. It gives a total number of station visits and stations are subsquently sorted descending by that value. Lillipop chart helps to visualize the mount of visits per station.

6 Calculation of Distance Between Stations

6.1 Distance calculation by using of FOR cycle

```
Number of bike rides, where start and end stations are identical is 65077.
Wall time: 18 s
```

The first approach calculates the distance with usage of method called “Spherical Law of Cosines.” For distance calculation python function is defined taking 4 parameters (geographical latitude and longitude for start and end station) as an input. FOR cycle iterates through all bike sharing entries, while all entries with identical start and end station are dropped. Parameter “%%time” is included to output the wall time for checking of algorithm time efficiency.

6.2 Distance calculation by using of LAMBDA function

```
Number of bike rides, where start and end stations are identical is 65077.
Wall time: 13 s
```

The second approach is based on the LAMBDA function, which seems to be less time consuming than FOR cycle and therefore more efficient.

7 Dataframe Preparation and Basic Statistics

Data frame preparation:

-> original dataframe from *.csv is joined with the new dataframe column containing a distance between stations for every single ride.

-> "Start date" as a string type is converted to datetime type, individual datetime items (Day, Month, Year, Time, Date) are segregated and saved into a special column for data aggregation option.

-> rows where distance is zero (start and end station are identical) are dropped and new dataframe is saved.

-> Iterquartile range Q3-Q1, "1.5 IQR above the third quartile", 25 % percentile, 75% percentile and Median are calculated for distance and duration column

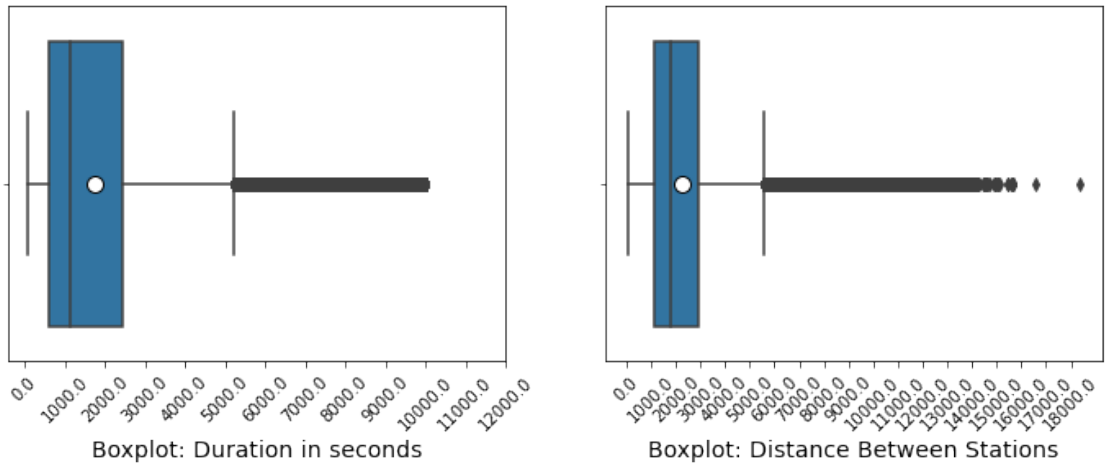
-> Data from duration and distance column are cut off by "1.5 IQR above the third quartile" value and later used for histogram plot.

```
Duration Data Column - 25% Percentile: 624.0
Duration Data Column - Median: 1163.0
Duration Data Column - 75% Percentile: 2529.0
Duration IQR: 1905.0
1.5 IQR above the third quartile: 5386.5
----------------------------------------------------------------------
----------------------------------------------------------------------
Distance Data Column - 25% Percentile: 1124.99
Distance Data Column - Median: 1817.1
Distance Data Column - 75% Percentile: 2904.61
Distance IQR: 1779.62
1.5 IQR above the third quartile: 5574.04
```

7.1 Identification of outliers: Rental Duration and Distance Between Stations

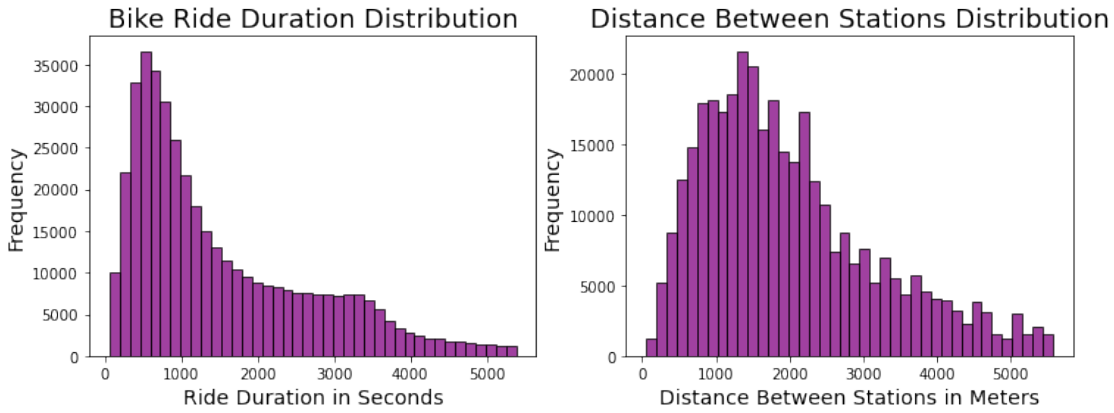
"duration" column : 1.5 IQR above the third quartile = 5386.5 s -> Boxplot is cut off by value of 10000 for better visualisation , but the outliers are still visible.

"Distance_in_meters" : 1.5 IQR above the third quartile = 5574.04 m -> Boxplot is cut off by value of 20000 for better visualisation, but the outliers are still visible.



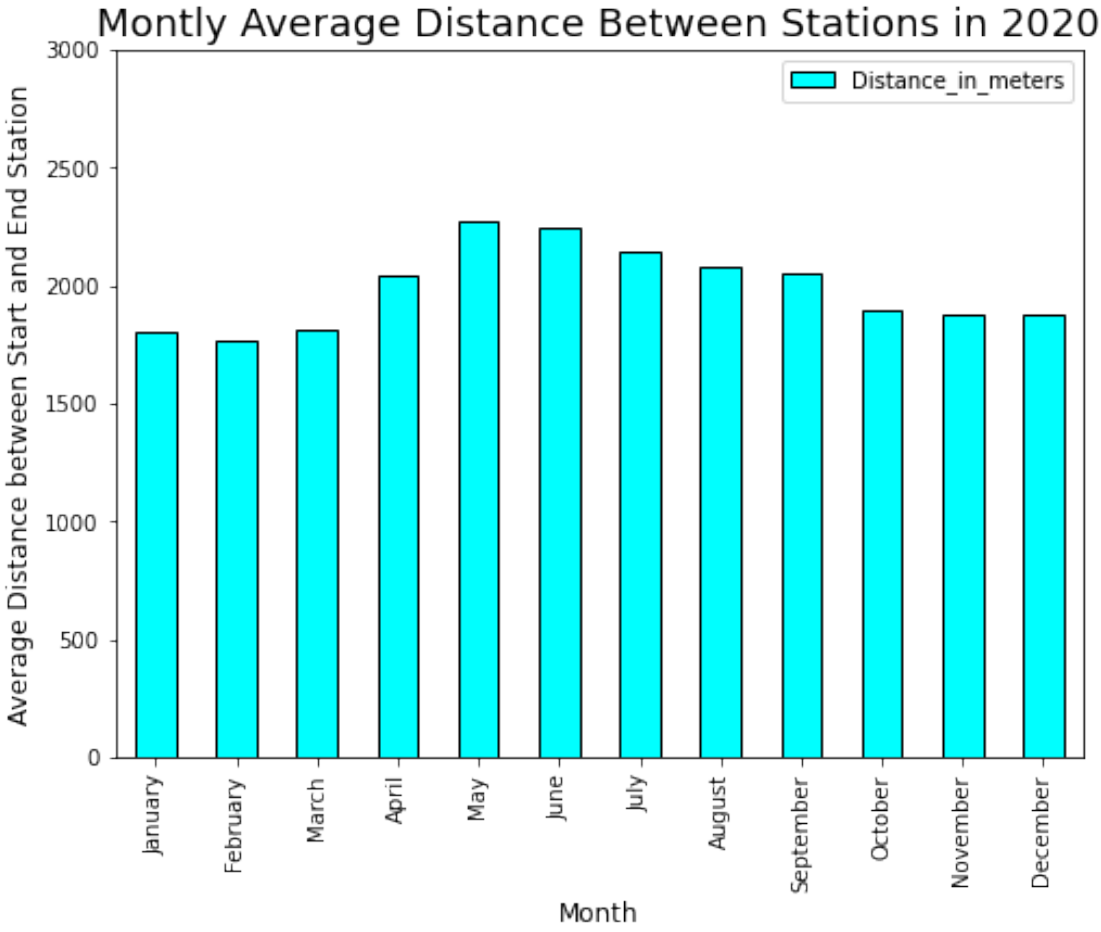
Box Plots are very good graphics for visualisation of data distribution, data quartiles and outlaying values.

7.2 Histograms: Rental Duration and Distance Between Stations



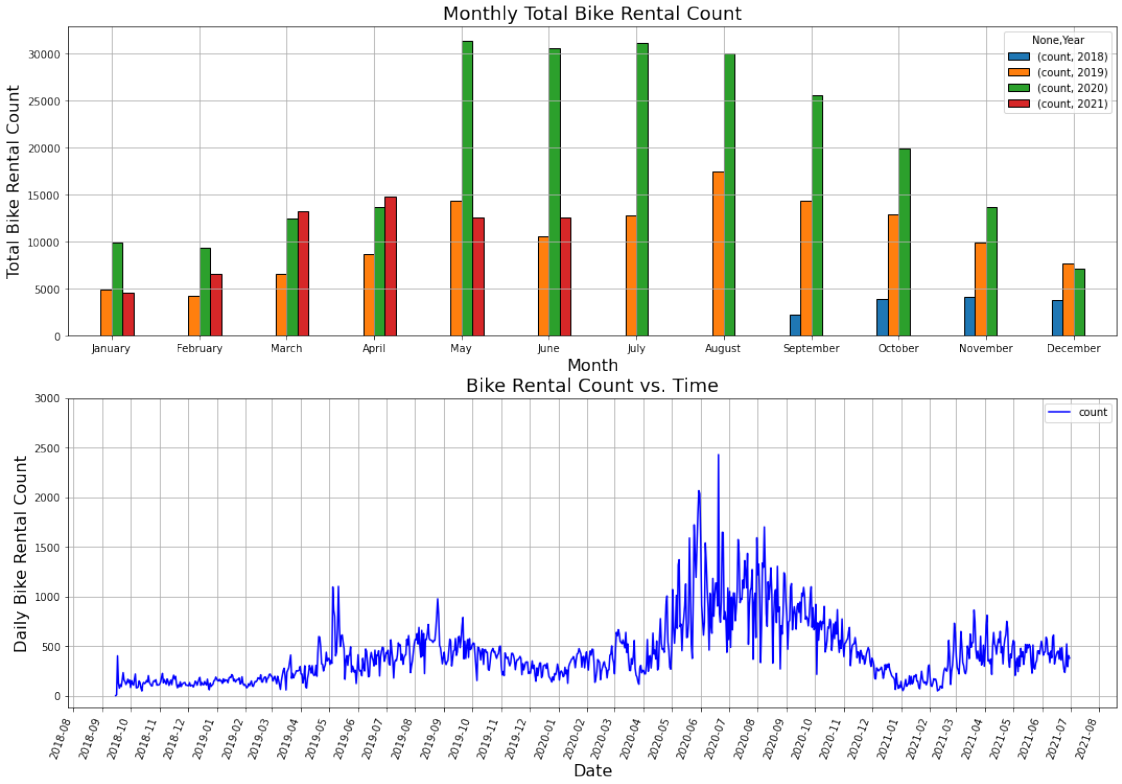
Histograms show the distribution of duration and distance data columns, while data points above value “1.5 IQR above the third quartile” are dropped.

7.3 How does the average distance between stations change throughout the year?



Bar chart shows how the average distance between stations for a single bike ride changes in every month in 2020. Average air distance between stations increases in a period between March and May. Starting from June an average distance between stations was monotonically decreasing.

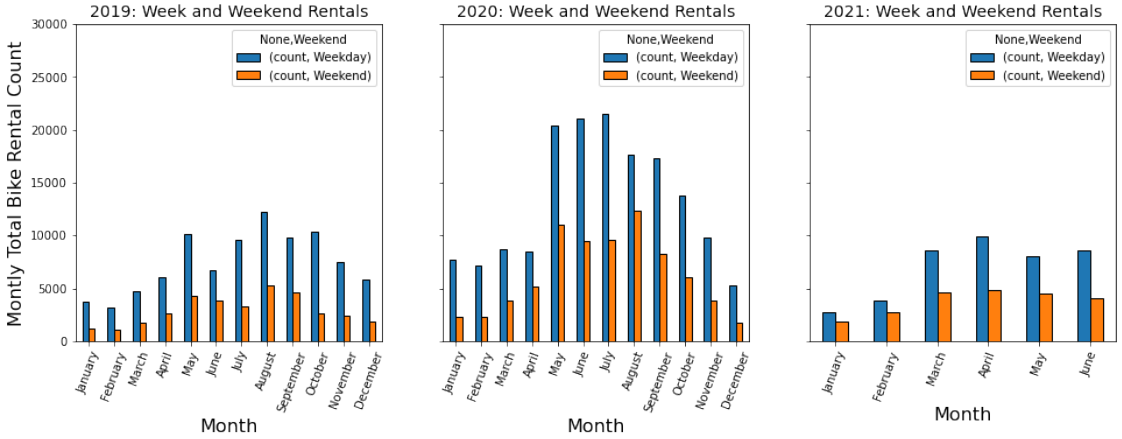
7.4 How does the bike rental count change in time?



Largest number of bike rentals is obvious in year 2020 and the values are doubled in comparison with 2019. Dropdown in 2021 can be clearly seen. Very profitable bike sharing business in 2020 was probably strongly supported by Covid-19 pandemic, during which people were less motivated to commute by public means of transport.

Very high demand in Summer 2020 seen on the bar chart correlates well with the demand oscillation seen on the time history line chart. Total counts are significantly oscillating in the period between May 2020 and October 2020. The reasons for oscillations of daily count can be further analyzed in detail.

7.5 Week and Weekend Amount of Bike Rentals for Each Month in 2020



Weekend bike rides mostly make up about half of the total bike rental per week.

==========

How to filter lines by 'date' column? One example approach below:

8 How does the weather condition influence the total number of bike rentals per day?

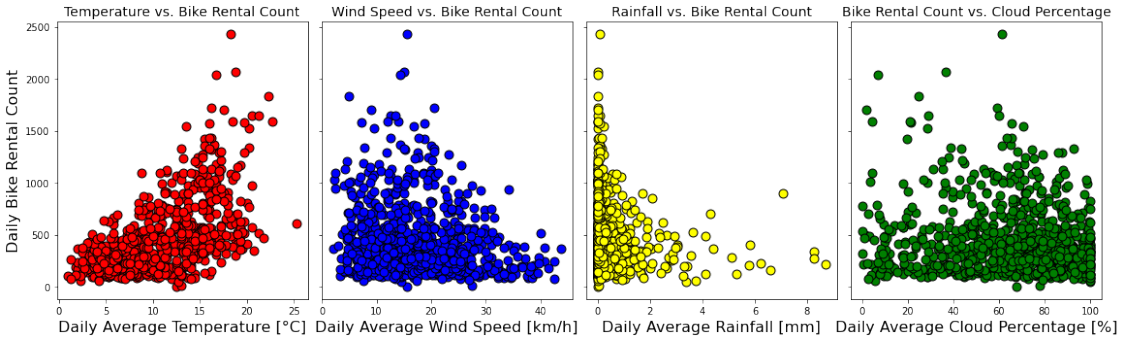
	time	Temp_in_Celsius	Wind_in_kmh	Rain_in_mm	Cloud_in_Percents	\
0	00:00	11	9	0.0	13	
1	03:00	13	11	0.0	96	
2	06:00	14	11	0.0	100	
3	09:00	14	14	0.1	78	
4	12:00	16	15	0.0	87	
5	15:00	17	14	0.0	97	
6	18:00	16	12	0.0	98	
7	21:00	14	11	0.0	52	
8	00:00	13	10	0.0	31	
9	03:00	13	8	0.0	52	

	date	Weekend
0	2018-09-01	Weekend
1	2018-09-01	Weekend
2	2018-09-01	Weekend
3	2018-09-01	Weekend
4	2018-09-01	Weekend
5	2018-09-01	Weekend
6	2018-09-01	Weekend
7	2018-09-01	Weekend

```
8  2018-09-02  Weekend
9  2018-09-02  Weekend
```

Table above contains the most important weather data saved as integer or float data type in order to do math operations. Weather data are available in the period from 01-09-2018 to 31-10-2020

8.1 Amount of Bike Rentals vs. Weather Condition: data from 15-09-2018 to 31-10-2020



Scatter plot shows the set of bivariate data. It shows how the bike rentals are influenced by temperature, wind, rainfall and cloud percentage. Every spot represents daily measurement. X-variable is explanatory variable and Y-variable is response variable. Other names for X and Y include the independent and dependent variables, respectively.

Left Window: Temperature and daily rental count have positive correlation, because data show uphill pattern moving from left to right. The number of bike rentals exceeds the value of 1000 only on days with the measured average temperature higher than 15°C.

Middle Left Window: Wind and daily bike rental seems to have negative correlation. Daily rental count higher than 1000 was achieved on days with the average wind speed not higher than 25 km/h.

Middle Right Window: Rainfall and daily bike rental seems to have negative correlation. For days with average rainfall higher than 2 mm is the daily bike rental count always below value of 1000.

Right Window: Correlation between variables is not clear from the scatter plot.

9 How does the weather condition correlate with the total number of bike rides per day

9.1 Detail view of Season With Maximal Amount of Bike Rentals: from 05-2020 to 09-2020

Detail view of time history data for bike rental count, daily average temperature, wind speed, rainfall and cloud percentage is show on the graphs below in the specific time period between 05-2020 and 09-2020. Since all graphs have synchronized X-axis the correlation between data can be also visually recognized on the line charts for a specific day.

Negative correlation between daily rental count vs. rainfall and cloud percentage can be seen on the line charts. Daily rental count is decreasing while the rainfall and cloud percentage are increasing at the same moment.

Since the scatter plot don´t adress the issue of whether or not the linear relationship was strong or weak, correlation coefficient calculation is better approach how to determine the correlation between columns. Pearson correlation coefficient is one of the measures how to judge the correlation between data columns. Outliers very far away from the mean value can significantly influence the correlation coefficient!

9.2 Pearson linear correlation

Pearson correlation coefficient is a measure of linear correlation between two sets of data. It is the ratio between the covariance of two variables and the product of their standard deviations; thus it is essentially a normalized measurement of the covariance, such that the result always has a value between −1 and 1.

Negative correlation = one variable increases, second decreases -> correlation is not dependent on the scale of data -> smaller p-value - more confidence we have in predictions

Correlation is always between -1 and +1:

-> perfect negative linear relationship = -1.00

-> strong negative relationship = -0.70

-> moderate negative relationship = -0.50

-> weak negative relationship = -0.30

-> no linear relationship = 0.00

-> weak positive relationship = +0.30

-> moderate positive relationship = +0.50

-> strong positive relationship = +0.70

-> perfect linear relationship = +1.00

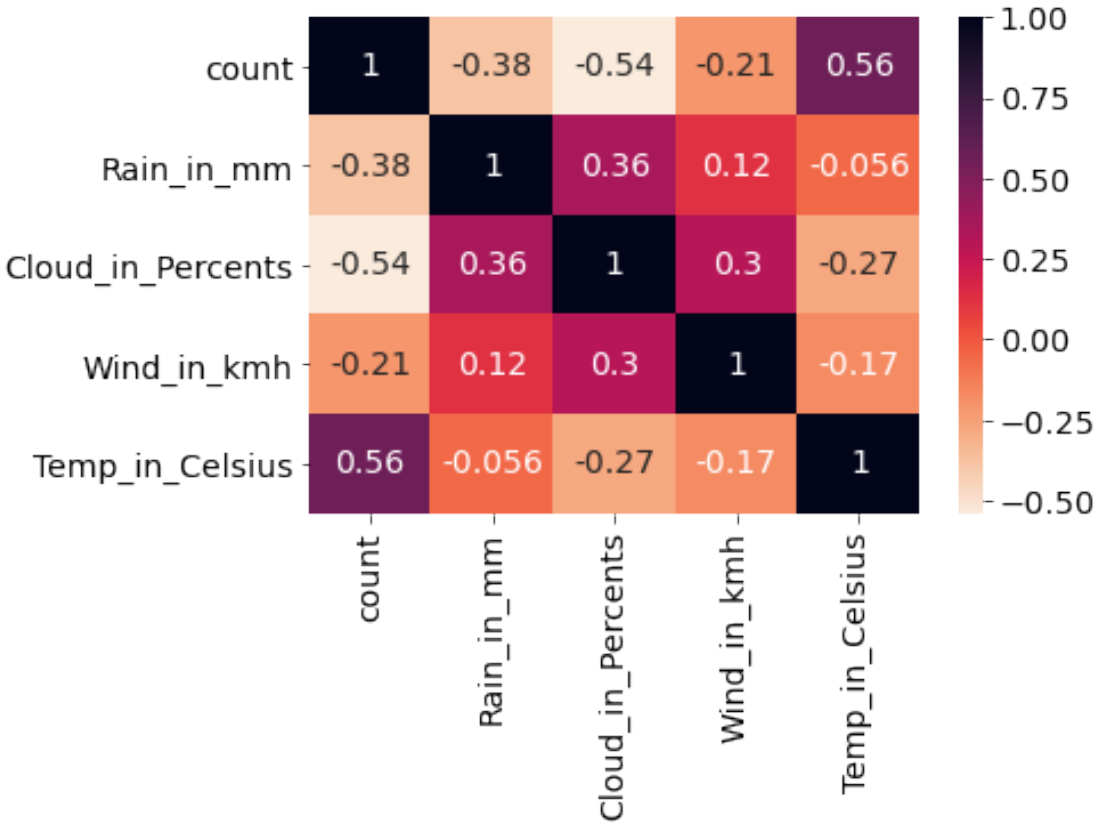
https://en.wikipedia.org/wiki/Pearson_correlation_coefficient

9.3 Pearson Linear Correlation Coefficients: data from 01-05-2020 to 30-09-2020

	count	Rain_in_mm	Cloud_in_Percents	Wind_in_kmh	\
count	1.000000	-0.382993	-0.540081	-0.208363	
Rain_in_mm	-0.382993	1.000000	0.360208	0.122370	
Cloud_in_Percents	-0.540081	0.360208	1.000000	0.303520	
Wind_in_kmh	-0.208363	0.122370	0.303520	1.000000	
Temp_in_Celsius	0.560968	-0.056087	-0.274930	-0.172508	

	Temp_in_Celsius
count	0.560968
Rain_in_mm	-0.056087
Cloud_in_Percents	-0.274930
Wind_in_kmh	-0.172508
Temp_in_Celsius	1.000000

9.4 Heatmap: Linear Correlation Coefficients Visualisation



Heatmap is better visualising the correlation coefficients between individual columns.

It confirms negative correlation between could percentage/rainfall and bike rental count, positive correlation is found between average daily temperature and the bike rental count.

9.5 Python Seaborn Data Plot and Regression Model: data from 01-05-2020 to 30-09-2020

lmplot in seaborn builts linear regression model including confidence bands. Bands represents all confidence intervals for every possible x and are tightest where data is grouped more densely.

In statistical modeling, regression analysis is a set of statistical processes for estimating the relationships between a dependent variable (often called the 'outcome' or 'response' variable) and one or more independent variables (often called 'predictors', 'covariates', 'explanatory variables' or 'features'). The most common form of regression analysis is linear regression, in which one finds the line (or a more complex linear combination) that most closely fits the data according to a specific mathematical criterion. For example, the method of ordinary least squares computes the unique line (or hyperplane) that minimizes the sum of squared differences between the true data and that line (or hyperplane).

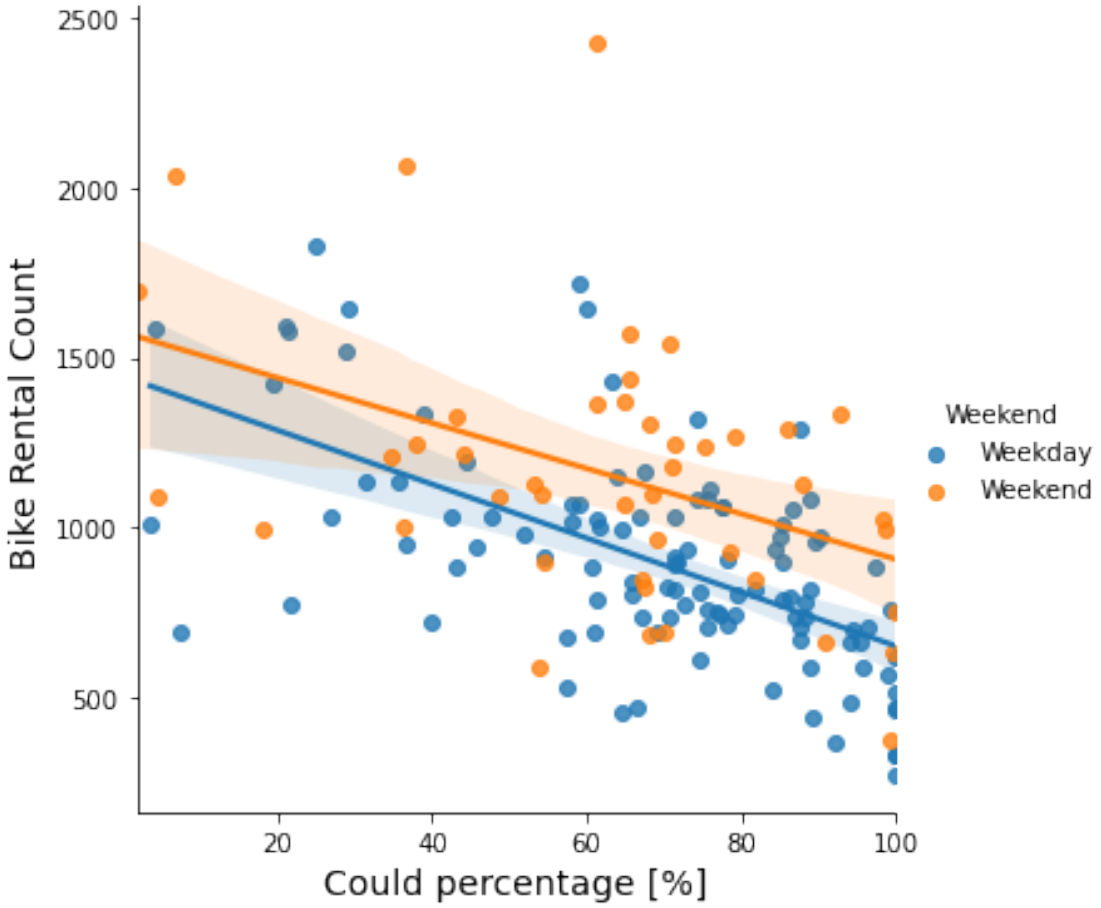
Regression attempts to establish how X causes Y to change. Correlation is a single statistic value,

whereas regression produces an entire equation.

https://en.wikipedia.org/wiki/Regression_analysis

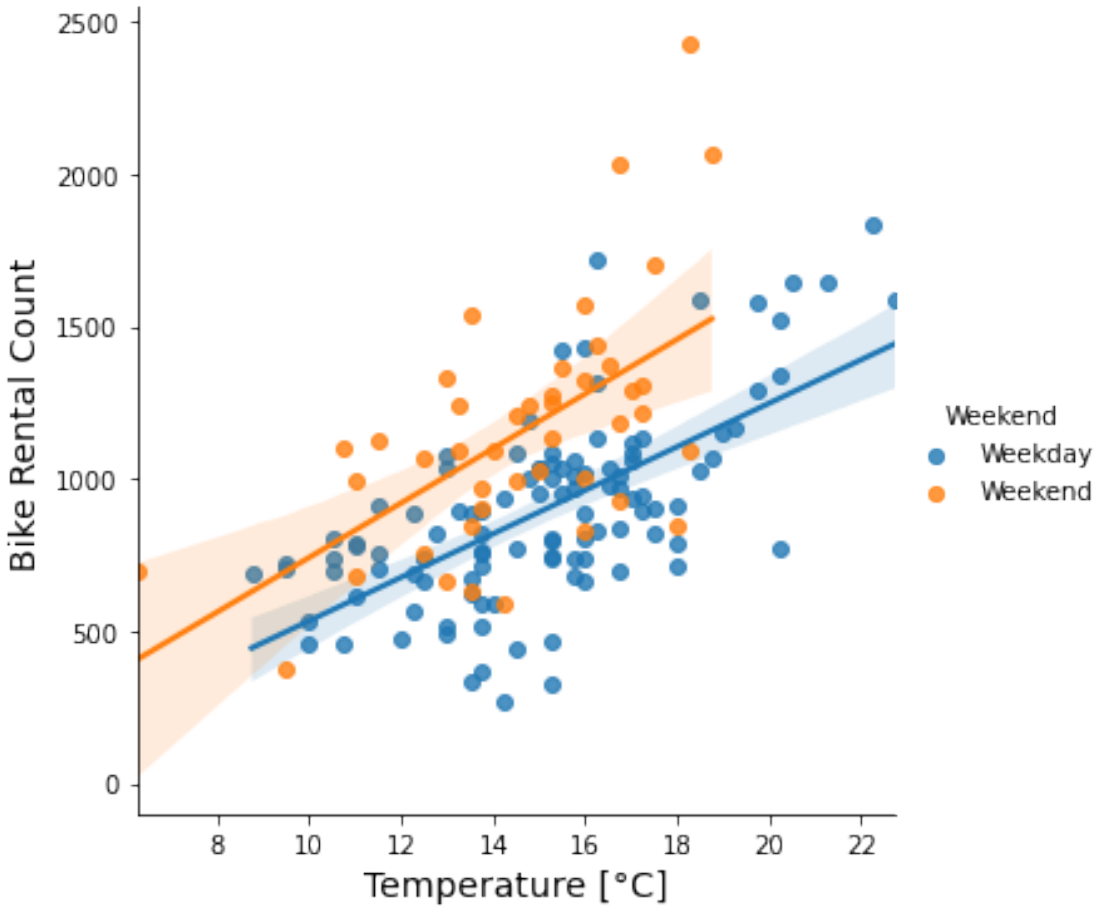
Moderate correlation has been established for temperature and cloud percentage variables and therefore regression analysis can follow.

```
Text(16.63270833333334, 0.5, 'Bike Rental Count')
```



Number of bike rentals tends to be higher on weekend days at the similar cloud percentage.

```
Text(16.63270833333334, 0.5, 'Bike Rental Count')
```



Number of bike rentals tends to be higher on weekend days at the similar temperature value.

Positive correlation between temperature and the rental count is seen on the regression line.

Would be interesting to ask what are the real reasons for the outliers. E.g. public holiday may significantly increase the daily rental count. Maybe there are more additional parameters that have nothing to do with weather condition and have a huge impact on the amount of bike rentals.

10 Conclusion

1. Picady Place, Royal Highland Show, Cycling Scotland Conference, Edinburgh University Sports Fair, Depot are 5 stations with the lowest number of visits and can be considered as inactive.
2. Portobello Kings Road, Victoria Quay, Duke Street, Leith Walk, Canonmills have highest bike surplus in the given time period from 09-2018 to 06-2021 in Edinburgh.
3. Bristo Square, St. Andrews Square, Charlotte Square, City Chambers, Pollock Halls have highest bike shortage in an given time period from 09-2018 to 06-2021 in Edinburgh.

4. Meadows East, Portobello Kings Road, Victoria Quay, Meadow Place, Canonmills are the busiest stations for bike sharing in Edinburgh.
5. Very short bike rides of duration of approximately 10-15 minutes are the most common. Bike rental for more than one day occured only ones in the data set and was identified as clear outlying value.
6. The most common air distance between stations is 0.5 - 1.5 km for a single ride.
7. Air distance between stations in 2020 was longest during spring and summer season. Average montly distance between stations for a bike ride was 2.0-2.5 km.
8. The highest increase of bike rides occured between April and May, 2020. Total montly count of rentals is in May more than 2 times higher than in April 2020.
9. The season from May 2020 to October 2020 shows largest monthly number of bike rentals.
10. Demand oscillation seen on the time history correlates well with conclusion point nr. 9.
11. Weekend bike rides make up about half of the total bike rental per week.
12. Strongest correlation between the number of bike rentals and weather was identified for temperature and cloud percentage parameter, but the correlation coefficient indicates a moderate correlation (around +/- 0.5). Correlation with the temperature is positive, correlation with cloud percentage is negative. Wind speed and rainfall are less important parameters influencing demand for bike sharing.
13. Outliers are influencing the linear correlation coefficient and may have more leverage and may change the scope of the regression line. Clear reasons for them should to be find and it is recommended to think about their exclusion from data set.