Project 1

Analyzing London Bike Share Data

Authors:
Yuzhen Ding (s196311)
Ömer Yılmaz (s196596)
Akmaral Pavlickova (s200605)

	Yuzhen	Ömer	Akmaral
Part 1	-	-	100%
Part 2	80%	-	20%
Part 3	40%	60 %	-
Part 4	33.3%	33.3%	33.3%

Part 1: A description of our data set

We will be working with London bike share data containing information on the number of short term rented bikes per hour (Santander Cycles) as well as information about the weather and bank holiday indicator in the period between 2015-01-02 and 2017-01-01.

Data can be downloaded here:

https://www.kaggle.com/hmavrodiev/london-bike-sharing-dataset

It contains data aggregated from 3 sources:

- Bike share data from: https://cycling.data.tfl.gov.uk/
- Weather data from: freemeteo.com
- UK bank holiday data from: https://www.gov.uk/bank-holidays

The raw dataset contains 17414 observations and contains information about the amount of bikes rented per hour, temperature (real and ‘feels like’), humidity, wind speed, weather code and weekend and bank holiday indicator. The attributes will be described in detail in Part 2.

Our main machine learning aim is to apply **regression** algorithm to predict the amount of bikes rented using the information about the time of the day, weather conditions and whether it was a weekend or a bank holiday. Before conducting an exploratory analysis of the dataset, we expect the time of the day being a strong predictor of the bikes rented (peak hours) as well as the weather conditions (‘good’ weather as mild temperature and no wind correlating positively with the amount of bikes rented).

We would like to see how well the model based on these attributes performs and is able to predict the amount of bikes rented.

This dataset is probably not the most suitable for **classification**, we would like to try to apply this technique nevertheless. Removing the date from the timestamp variable, a possible application of this technique would be predicting whether it was weekend or not (binary classification) as we expect that the amount of bikes rented during the day won't be following the same shape during the weekdays (increasing during peak hours in the morning and in the late afternoon) and during the weekends (the shape being more flat).

Our dataset is very poorly suited for applying **association mining** as there are no rules to be discovered with the given data. If the dataset would contain specific user data (instead of aggregated amount of bikes rented per hour) that would include whether the user uses bike share services, other public transportation (metro, overground, bus, train), car share services or taxi-like services (uber, taxi), it would be interesting to apply this technique to try to discover if taking a specific mode of transportation (for example taking an uber ride) is associated with users being more likely to use the bike sharing services.

We believe that **clustering** cannot be applied easily either. If the dataset would contain user data with additional information about the user together to their history usage of bike sharing service, we could potentially apply this method to discover groups of users (clusters) from the data without labels.

We could potentially apply **anomaly detection** algorithms and try to detect outliers for the bike shares. It would be interesting to see if there are some anomalies in the dataset where the number of bikes rented deviates significantly from other observation given the time of the day, temperature, being it weekend or holiday. That could be due to a 'natural' event like a bike concert or another cultural event on a Tuesday that would result in a spike of bikes rented in the late evening hours. Or it could simply be a wrong signal/bug in the bike sharing station that sent the wrong message to the main database/cloud.

Apart from exploratory analysis, we did not find any other analysis using the dataset that would be publicly available.

Part 2: A detailed explanation of the attributes of the data

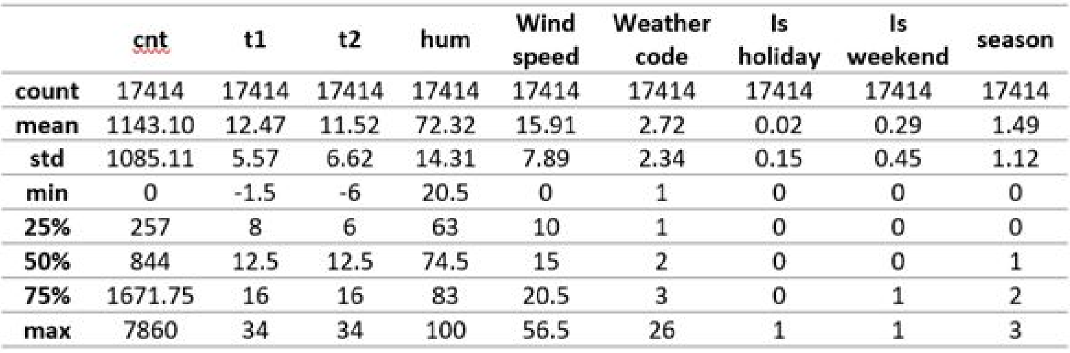
The raw dataset contains the following attributes:

Attribute	Description	Possible values	Classification
timestamp	Datetime		Discrete, interval
cnt	Count of new bike shares		Discrete, ratio
t1	Real temperature in Celsius		Continuous, interval
t2	'Feels like' temperature in Celsius		Continuous, interval

hum	Humidity in %		Continuous, interval
wind_speed	Wind speed in km/h		Continuous, ratio
Weather_ code	Weather category	1 = Clear; mostly clear but have some values with haze/fog/patches of fog/ fog in vicinity 2 = Scattered clouds / few clouds 3 = Broken clouds 4 = Cloudy 7 = Rain/ light rain shower/ light rain 10 = Rain with thunderstorm 26 = Snowfall 94 = Freezing fog	Discrete, nominal
is_holiday	Bank holiday indicator	1 = UK bank holiday 0 = non holiday	Discrete, nominal
is_weekend	Weekend indicator	1 = weekend 0 = weekday	Discrete, nominal
season	Season category	0 = spring 1 = summer 2 = fall 3 = winter	Discrete, nominal

For detecting **missing values** in Python, function X.isnull().sum() is used where X represents raw data. The result is 0 missing values.

The basic **summary** statistics of the attributes are shown below.



	cnt	t1	t2	hum	Wind speed	Weather code	Is holiday	Is weekend	season
count	17414	17414	17414	17414	17414	17414	17414	17414	17414
mean	1143.10	12.47	11.52	72.32	15.91	2.72	0.02	0.29	1.49
std	1085.11	5.57	6.62	14.31	7.89	2.34	0.15	0.45	1.12
min	0	-1.5	-6	20.5	0	1	0	0	0
25%	257	8	6	63	10	1	0	0	0
50%	844	12.5	12.5	74.5	15	2	0	0	1
75%	1671.75	16	16	83	20.5	3	0	1	2
max	7860	34	34	100	56.5	26	1	1	3

From the summary, the data size is 17414, which records the number of people renting shared bikes in every hour for two years from 4/01/2015 to 3/01/2017 in London.

In peak hours, the maximum count of new bike shares up to 7860, while in the slack period, especially late at night, nobody is willing to rent a bike.

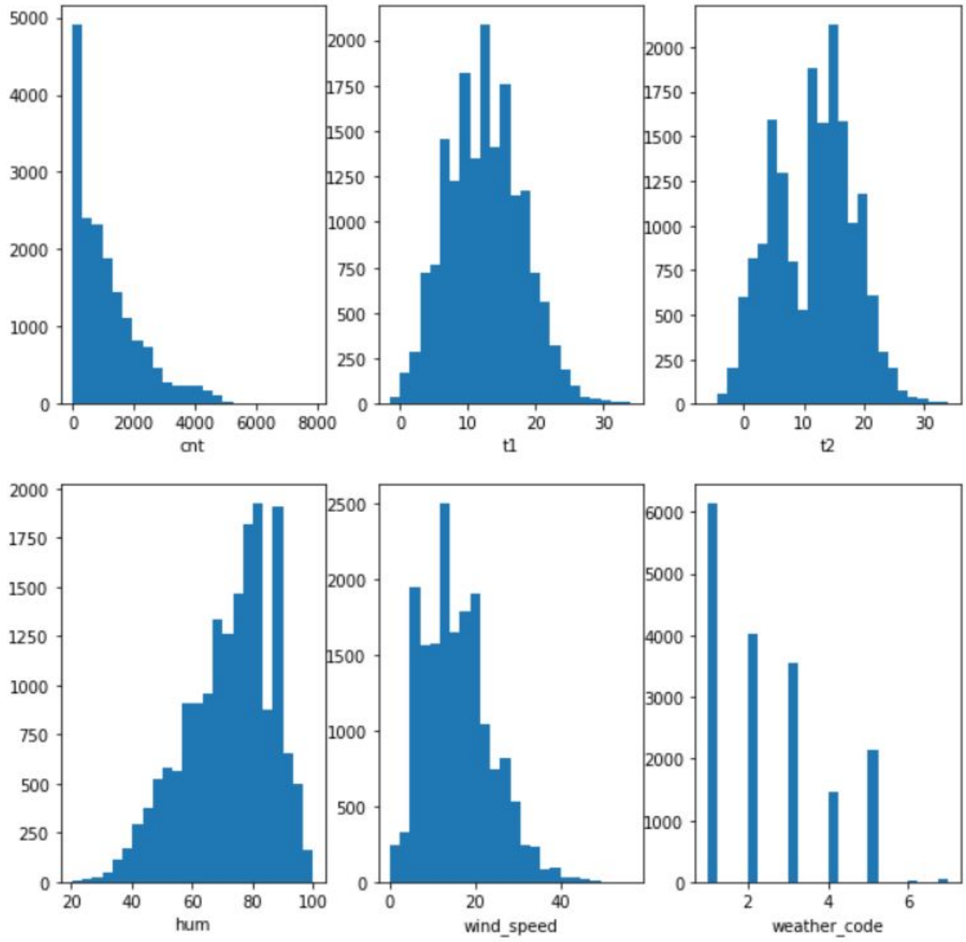
Temperature is ranging from -1.5 to 34 degrees; likewise, during these two years, London is mostly humid with an average of 72.32.

Apart from that, wind speed, weather and season data are documented.

Lastly, whether it is holiday and weekend or not also affects the count of bike renting.

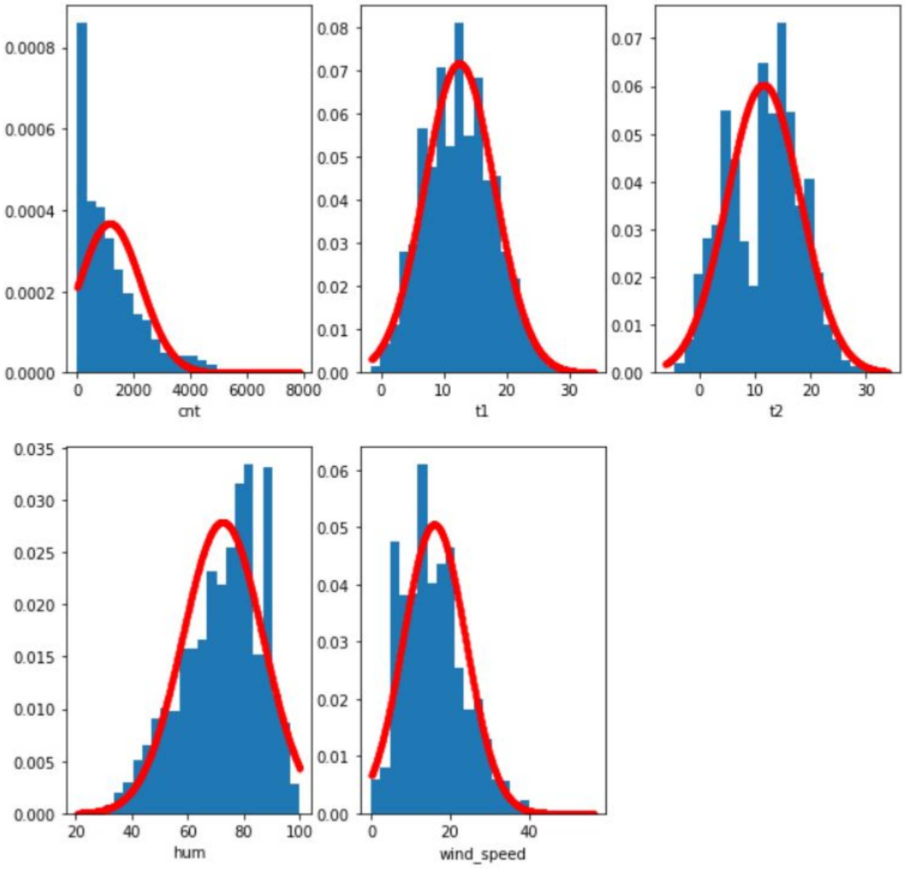
Part 3: Data visualization(s) based on suitable visualization techniques including a principal component analysis (PCA)

1) Histogram of Data



We can see the numbers of the attributes above. Since the other variables are only about time, distribution of them is uniform.

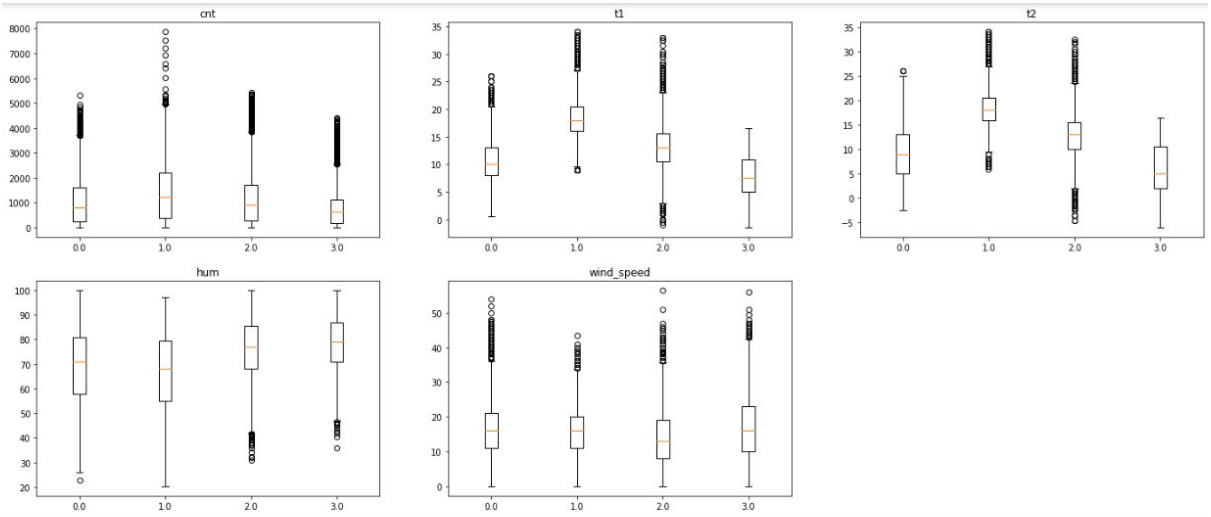
2) Comparison with Normal Distribution



When we compare the histograms to normal distribution with the same mean and variance, we get the visuals above. Attributes except cnt are nearly normal distributed. Cnt is not nearly normal distributed because the 0-333 block is huge.

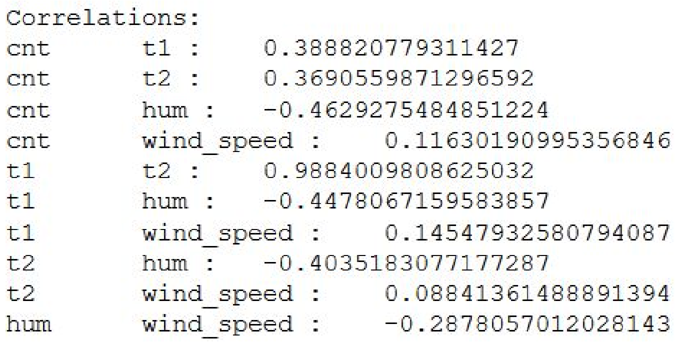
On the other hand, the other attributes are categorical or about time, it doesn't make any sense to compare them to normal distribution.

3) Outliers

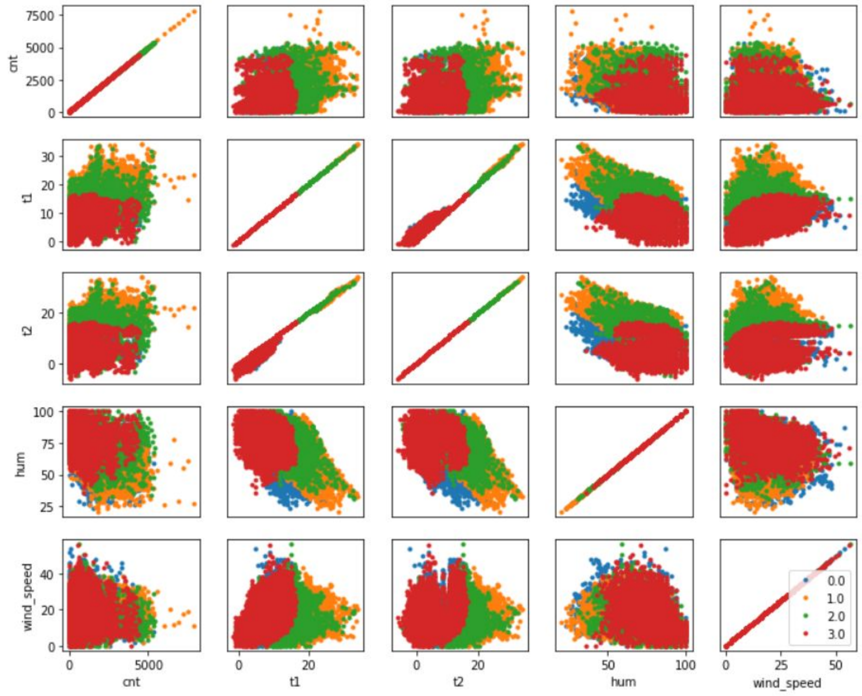


There are some outliers of course. The reason for the outliers would be some special events or days or extreme weather conditions. However, we have more than 17000 data, so outliers are not that much. *(Box plots contain seasons on horizontal axis).*

4) Correlations



```
Correlations:
cnt      t1 :    0.388820779311427
cnt      t2 :    0.3690559871296592
cnt      hum :   -0.4629275484851224
cnt      wind_speed :    0.11630190995356846
t1       t2 :    0.9884009808625032
t1       hum :   -0.4478067159583857
t1       wind_speed :    0.14547932580794087
t2       hum :   -0.4035183077177287
t2       wind_speed :    0.08841361488891394
hum      wind_speed :    -0.2878057012028143
```



When we check the correlations between the attributes, t1 and t2 are perfectly correlated since one of them is apparent temperature and the other is real temperature. Also we can see the correlation between hum and t1. The reason for that is when temperature goes up, hum goes down. However, they are not perfectly correlated. *(Observations grouped by seasons, spring = 0, summer = 1, fall = 2, winter = 3).*

PCA section

1) Data manipulation

Before carrying out PCA, it is necessary to manipulate the data.

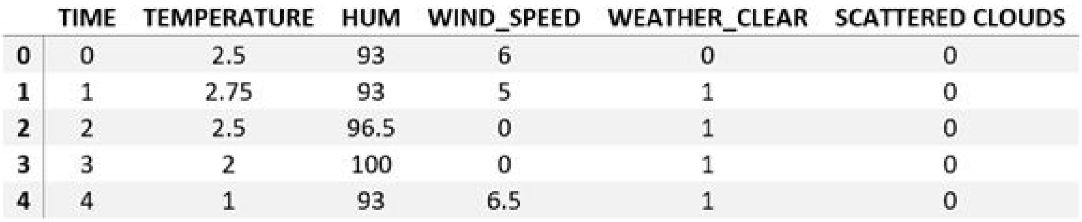
Firstly, the format of attribute timestamp is like ‘4/01/2015 0:00’ combining the date and time together. For analysing the influence of different dates and time, here we convert attribute timestamp into two separate columns, namely, attribute dates and attribute time.

Secondly, since the attribute t1 and t2 all indicate the temperature, we use a single representative feature temperature by taking the mean of t1 and t2.

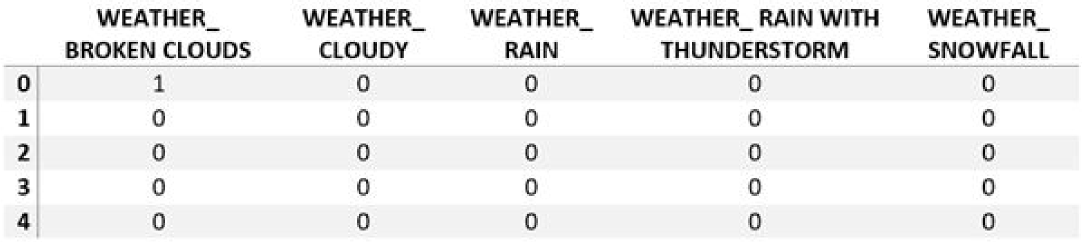
Thirdly, the attribute weather_code contains 8 different weather conditions, we utilise one-out-of-k encoding to deal with it. In fact, there is no weather_code =94 in the dataset. In addition, the attribute season is a categorical variable with values from 0 to 3 to indicate spring, summer, autumn and winter, so one-out-of-K encoding also applied in this variable.

Finally, create a new data frame X_new to store manipulated data with 17 columns including attributes such as humility, wind_speed and so on. The attribute cnt in the dataset is not included in X_new, because other attributes can highly influence the values of cnt. Besides, the date is excluded in X_new as which is not suitable for PCA.

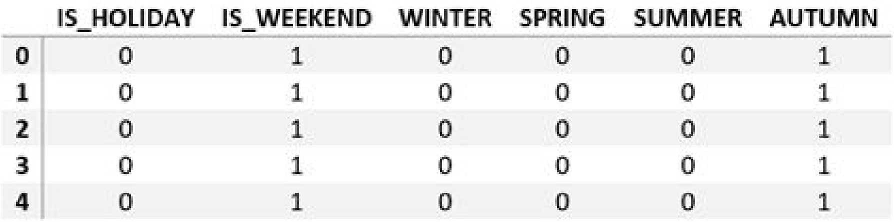
The table below shows the first 5 rows of the X_new with 17 attributes.



	TIME	TEMPERATURE	HUM	WIND_SPEED	WEATHER_CLEAR	SCATTERED CLOUDS
0	0	2.5	93	6	0	0
1	1	2.75	93	5	1	0
2	2	2.5	96.5	0	1	0
3	3	2	100	0	1	0
4	4	1	93	6.5	1	0

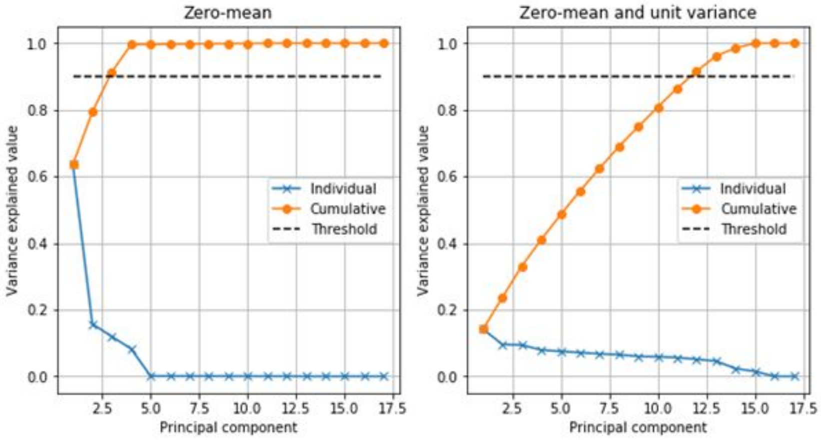


	WEATHER_ BROKEN CLOUDS	WEATHER_ CLOUDY	WEATHER_ RAIN	WEATHER_ RAIN WITH THUNDERSTORM	WEATHER_ SNOWFALL
0	1	0	0	0	0
1	0	0	0	0	0
2	0	0	0	0	0
3	0	0	0	0	0
4	0	0	0	0	0



	IS_HOLIDAY	IS_WEEKEND	WINTER	SPRING	SUMMER	AUTUMN
0	0	1	0	0	0	1
1	0	1	0	0	0	1
2	0	1	0	0	0	1
3	0	1	0	0	0	1
4	0	1	0	0	0	1

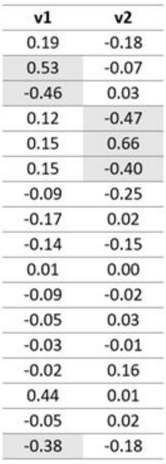
2) variation explanation



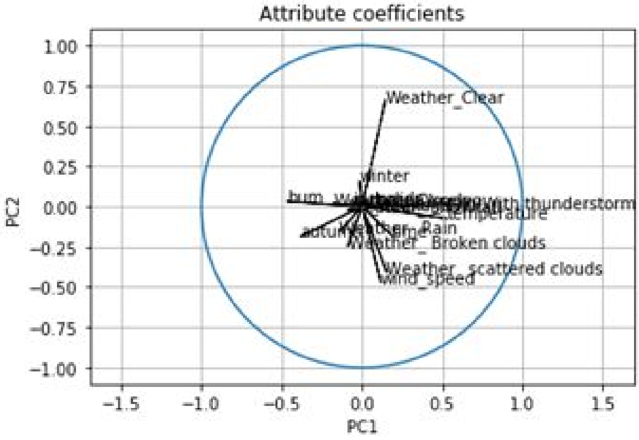
For applying PCA, centred data is taken. The left image above shows the ratio of key PCA components and its cumulative curve, which changes greatly after the X_new data being standardised as shown at the right image.

When data is only subtracted by its mean, PCA can efficiently use the first three components to represent the total 17 components as their cumulative ratio is higher than 90%. However, In the dataset X_new, attributes humidity and wind_speed have larger scales than others' which attributes to the relatively high ratio. Therefore, the data is standardized by the standard deviation prior to the PCA analysis. As a result, the figure on the right shows that PCA uses 12 components to represent the data.

3) the principal directions of the considered PCA components

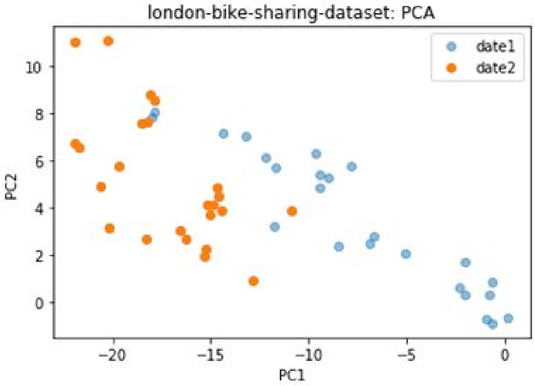


v1	v2
0.19	-0.18
0.53	-0.07
-0.46	0.03
0.12	-0.47
0.15	0.66
0.15	-0.40
-0.09	-0.25
-0.17	0.02
-0.14	-0.15
0.01	0.00
-0.09	-0.02
-0.05	0.03
-0.03	-0.01
-0.02	0.16
0.44	0.01
-0.05	0.02
-0.38	-0.18



The table is partially from the principal directions V after applying singular value decomposition (SVD) to the new data X_new. V1 emphasizes the component temperature, humidity and season in autumn, while V2 emphasizes wind_speed as well as the weather in clear and scattered clouds.

4) data projection



The figure shows the projection of the centred data onto the principal component space. The data is from the first 2 days, namely 4/01/2015 and 5/01/2015.

Part 4: What we have learned from data

After conducting the exploratory analysis it seems viable to apply a regression method to predict the amount of bike shares. ‘Real’ and ‘feels like’ temperature correlates nearly perfectly with each other, making it a good candidate for transformation or keeping only one of these attributes.

To apply classification to predict in which season the bike was rented seems also feasible given the variation of nearly all the attributes w.r.t the seasons.