Multivariate Exploratory Data Analysis

Shizhi Chen-10307389

Dataset description

The 'bike-sharing' dataset is downloaded from the UCI Machine Learning. Original Source:

Fanaee-T, Hadi, and Gama, Joao, 'Event labeling combining ensemble detectors and background knowledge', Progress in Artificial Intelligence (2013): pp. 1-15, Springer Berlin Heidelberg.

https://archive.ics.uci.edu/ml/datasets/Bike+Sharing+Dataset#

This dataset contains the daily count of rental bikes of 2012 in Capital bikeshare system with the corresponding weather and seasonal information.

Dataset Attribute Information:

- instant: record index
- season : spring, summer, fall, winter
- mnth : month (1 to 12)
- weekday : day of the week
- workingday : if day is neither weekend nor holiday is 1, otherwise is 0.
- weathersit :
- 1: Clear, Few clouds, Partly cloudy, Partly cloudy
- 2: Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist
- 3: Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain +

Scattered clouds

- 4: Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog

- temp: Normalized temperature in Celsius.

- hum: Normalized humidity.

- windspeed: Normalized wind speed.

- cnt: count of total rental bikes including both casual and registered

Dataset Quality

The rental counts data are collected automatically by the company back-end data system so there are no missing values and the data are reasonably accurate. However, the weather information in this dataset is collected by open source website, it may not be expected to be completely accurate. Therefore, I think the quality of this dataset is 3☆.

Exploratory Data Analysis

#Load the libraries needed

```
%matplotlib inline
import pandas as pd
import numpy as np
import scipy.stats as st
import matplotlib.pyplot as plt
import seaborn as sns
import statsmodels.formula.api as sm
sns.set_style("whitegrid", {'axes.grid' : False})
from pylab import *
import matplotlib.pyplot as plt
```

#Read the dataset

The first lines look like:

```
data = pd.read_csv('./bike_sharing_day.csv')
data.head()
```

	instant	dteday	season	yr	mnth	holiday	weekday	workingday	weathersit	temp	atemp	hum	windspeed	casual	registered	cnt
0	1	2012/1/1	winter	1	1	0	0	0	1	0.370000	0.375621	0.692500	0.192167	686	1608	2294
1	2	2012/1/2	winter	1	1	1	1	0	1	0.273043	0.252304	0.381304	0.329665	244	1707	1951
2	3	2012/1/3	winter	1	1	0	2	1	1	0.150000	0.126275	0.441250	0.365671	89	2147	2236
3	4	2012/1/4	winter	1	1	0	3	1	2	0.107500	0.119337	0.414583	0.184700	95	2273	2368
4	5	2012/1/5	winter	1	1	0	4	1	1	0.265833	0.278412	0.524167	0.129987	140	3132	3272

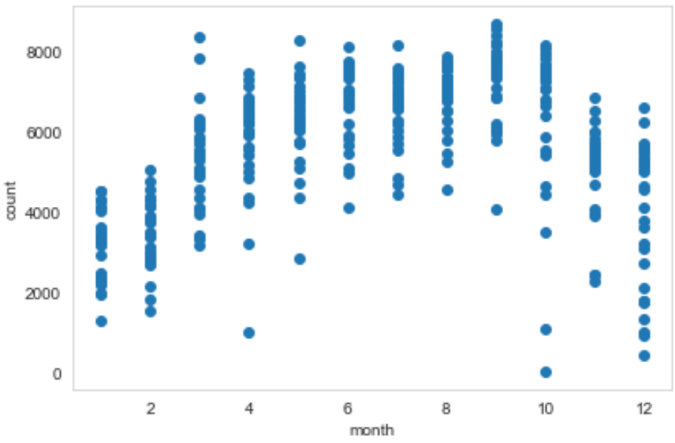
In this dataset, we are interested in what factors affect the number of times a bicycle is rented daily, so I will do some exploratory research on the relationship between different variables and cnt.

Select the variables we want to analyse

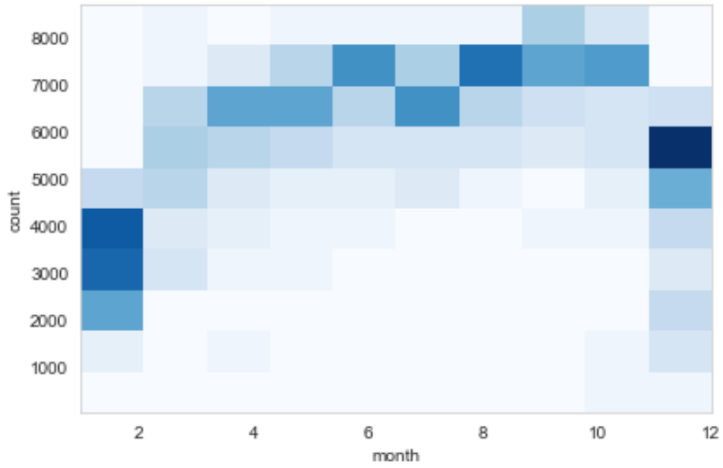
```
cnt=data['cnt']
mnth=data['mnth']
temp=data['temp']
hum=data['hum']
windspeed=data['windspeed']
weathersit=data['weathersit']
season=data['season']
```

First, we want to find if the count of rented sharing-bike is related to the time, so visualised them by some plots and 2d histogram

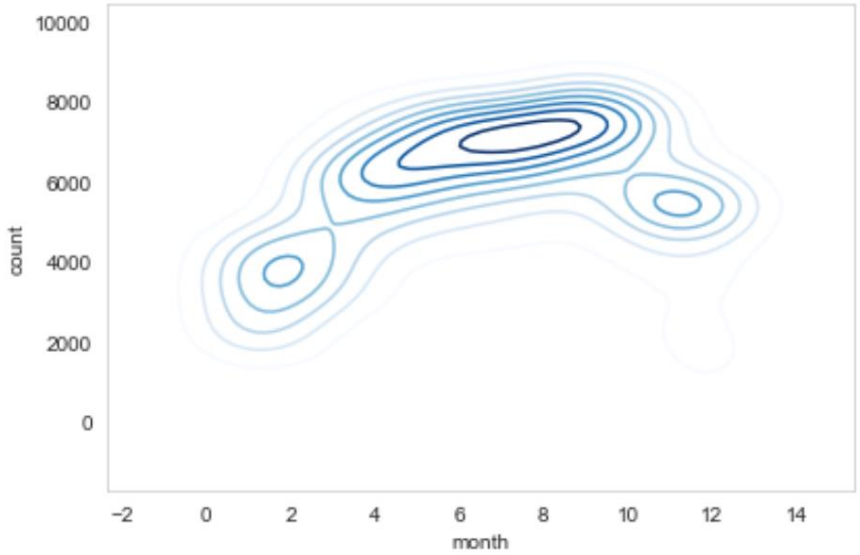
```
plt.figure(figsize=(6,4))
plt.scatter(mnth,cnt)
plt.xlabel('month')
plt.ylabel('count')
plt.tight_layout()
```



```
plt.figure(figsize=(6,4))
plt.hist2d(mnth,cnt,cmap='Blues')
plt.xlabel('month')
plt.ylabel('count')
plt.tight_layout()
```



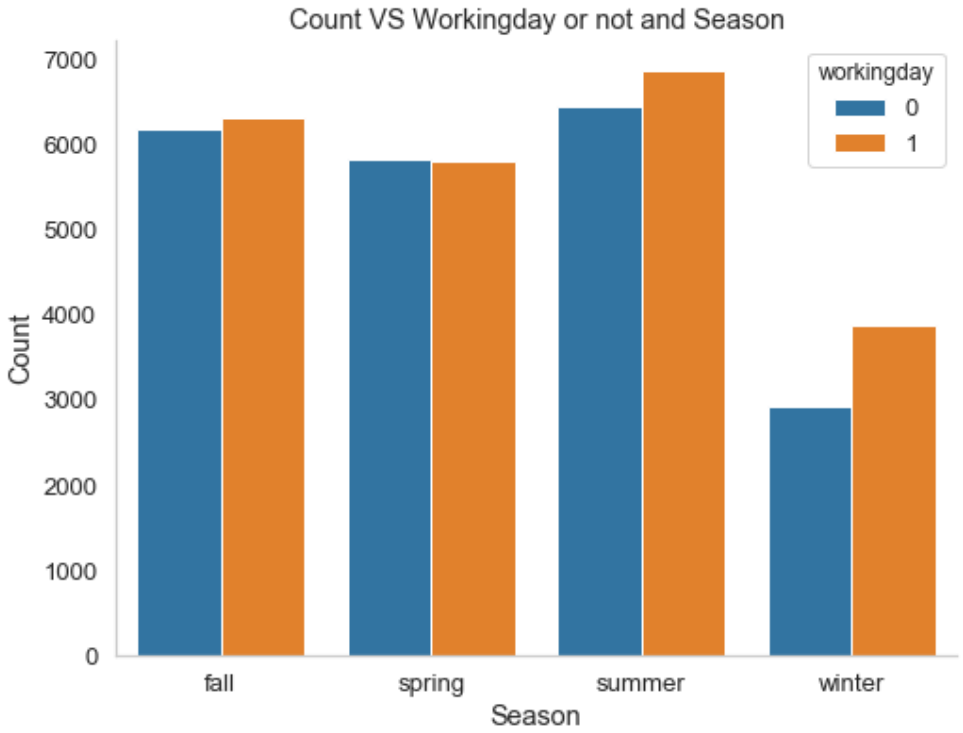

```
plt.figure(figsize=(6,4))
sns.kdeplot(mnth,cnt,cmap="Blues")
plt.xlabel('month')
plt.ylabel('count')
plt.tight_layout()
```



From the above figures, we can clearly see that the different time periods of one year have a greater impact on the use of shared bicycles. Therefore, we further analyse whether the use of shared bicycles is related to the season and whether it is a working day or not.

visualized them in a histogram

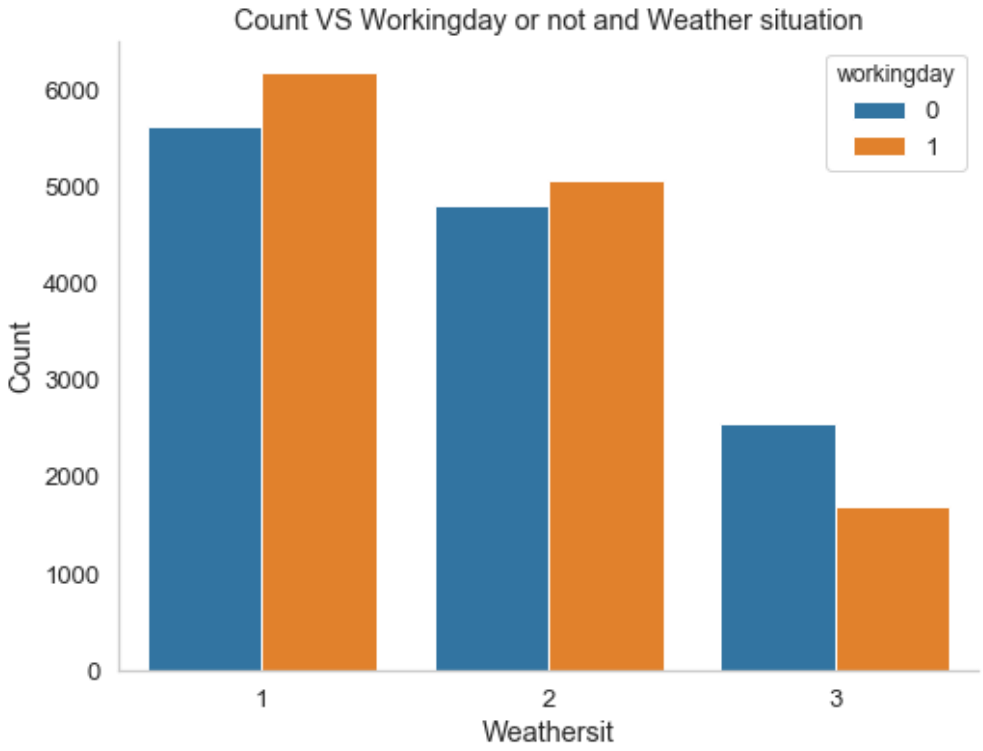
```
group1 = data.groupby(['season', 'workingday'])['cnt'].mean()
group1= pd.DataFrame(group1).reset_index()
#group_pivot = group1.pivot(index='season',columns='workingday',values='cnt')
plt.figure(figsize=(8, 6))
sns.set_context("notebook", font_scale=1.2, rc={"lines.linewidth": 2.5})
sboxplot2 = sns.barplot(x="season", y="cnt",hue='workingday', data=group1)
sns.despine(top=True)
plt.xlabel('Season')
plt.ylabel('Count')
plt.title('Count VS Workingday or not and Season')
plt.show()
```



As can be seen from the figure, the number of sharing-bike rented in winter is the lowest, the summer is the largest, and the number of people in spring and autumn is roughly equal. At the same time, people prefer to use sharing-bike on weekdays.

Based on the above analysis, the usage of shared bicycles is likely to be related to the weather to a certain extent. Therefore, analyse the weather and count is necessary.

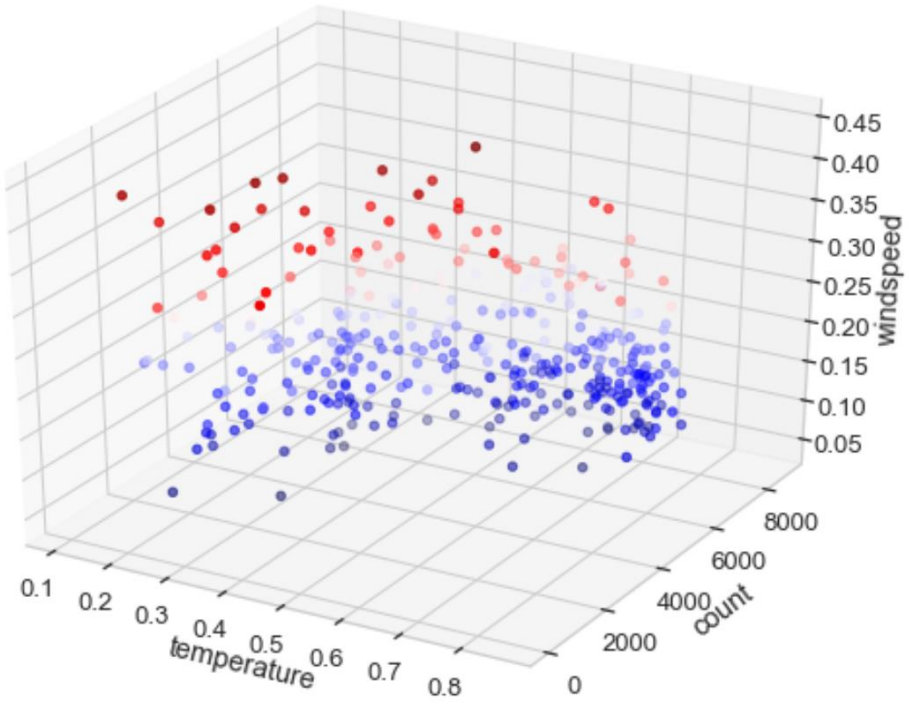
```
group1 = data.groupby(['weathersit', 'workingday'])['cnt'].mean()
group1= pd.DataFrame(group1).reset_index()
#group_pivot = group1.pivot(index='mnth',columns='workingday',values='cnt')
plt.figure(figsize=(8,6))
sns.set_context("notebook", font_scale=1.2, rc={"lines.linewidth": 2.5})
sboxplot2 = sns.barplot(x="weathersit", y="cnt",hue='workingday', data=group1)
sns.despine(top=True)
plt.xlabel('Weathersit')
plt.ylabel('Count')
plt.title('Count VS Workingday or not and Weather situation')
plt.show()
```



Obviously, the worse the weather, the fewer times the shared bicycle is used. Only during extreme weather periods, the number of people using shared bicycles on weekdays is lower than non-working days.

To be specific, two main weather situation factors that might influence the count of rental sharing-bike is temperature and wind speed. Thus, the relationship between the count, temperature and wind speed also should be analysed.

```
from mpl_toolkits.mplot3d import Axes3D
fig = plt.figure(figsize=(8,6))
ax = fig.add_subplot(111, projection='3d')
ax.scatter3D(temp,cnt,windspeed, marker='o', c=windspeed,cmap='seismic')
ax.set_xlabel('temp')
ax.set_ylabel('cnt')
ax.set_zlabel('windspeed')
plt.tight_layout()
```



The 3d Scatter shows that the points with large counts are concentrated at higher temperatures and lower wind speeds.

Next, in order to build a suitable model, analysing the relationship between various numerical variables is necessary.

```
#set an appropriate precission
#Turn off scientific counting method
```

```
np.set_printoptions(precision=4)
np.set_printoptions(suppress=True)
```

```
X=[temp,hum,windspeed,cnt]
```

```
xbar = np.mean(X,1)
print(xbar)
```

```
[    0.5041     0.6122     0.1896  5599.9344]
```

```
S = np.cov(X)
print(S)
```

```
[[       0.031        0.0028      -0.0028     224.8492]
 [       0.0028       0.018       -0.0031     -21.331 ]
 [      -0.0028      -0.0031       0.0061     -39.0228]
 [     224.8492     -21.331      -39.0228 3199332.7409]]
```

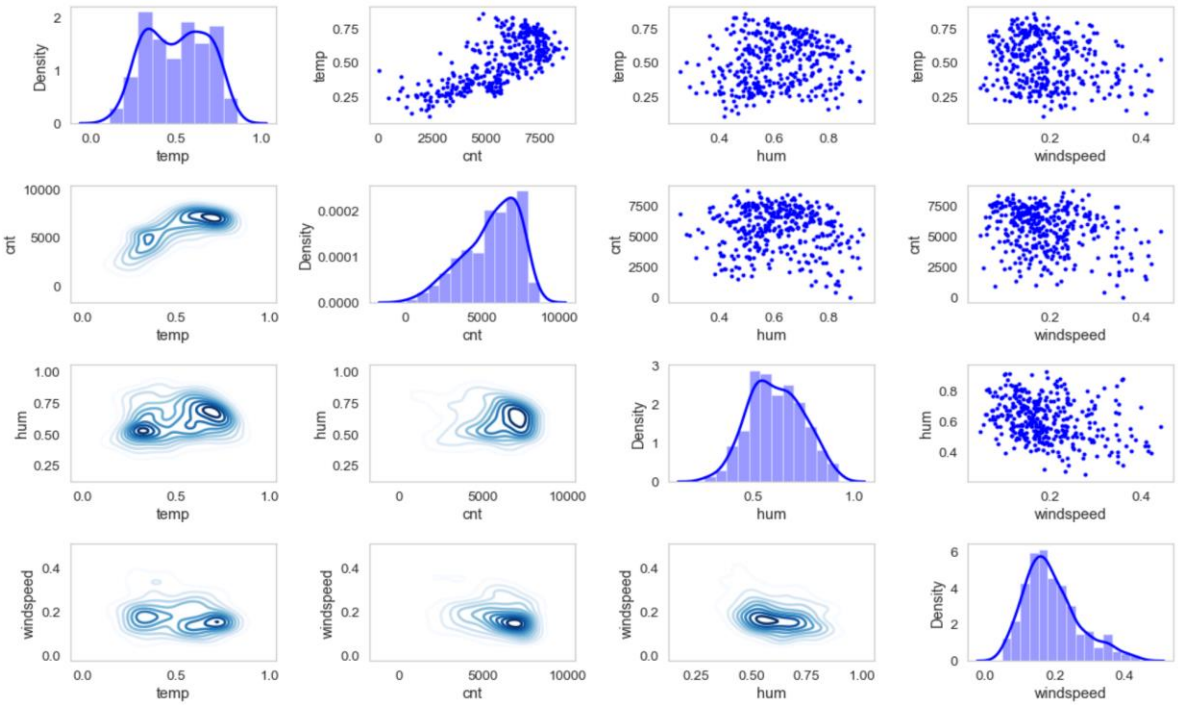
```
R = np.corrcoef(X)
print(R)
```

```
[[ 1.      0.1183 -0.2037  0.7138]
 [ 0.1183  1.     -0.291  -0.0889]
 [-0.2037 -0.291   1.     -0.279 ]
 [ 0.7138 -0.0889 -0.279   1.    ]]
```

```
names = ['temp','cnt','hum','windspeed']
```

```
X = np.array(X)
X = X.T
```

```
plt.figure(figsize=(16,10))
for i in range(0,4):
    for j in range(0,4):
        plt.subplot(4,4,1+i+(4*j))
        if i==j:
            sns.distplot(X[:,i],color="b")
            plt.xlabel(names[i])
            plt.ylabel('Density')
        else:
            if i<j:
                sns.kdeplot(np.ravel(X[:,i]),np.ravel(X[:,j]),cmap="Blues")
            else:
                plt.scatter(X[:,i],X[:,j],c="b",marker=".")
            plt.ylabel(names[j])
            plt.xlabel(names[i])
plt.tight_layout()
```



As can be seen from the results of all the above EDA, temperature, humidity, wind speed, working day and weather situation are considered highly related to the count of total rental sharing-bikes per day. Therefore, multiple linear regression models can be established between them.

```
model = sm.ols(formula="cnt ~ temp + hum + windspeed + weathersit + workingday", data=data).fit()
print(model.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                    cnt   R-squared:                       0.603
Model:                            OLS   Adj. R-squared:                  0.597
Method:                 Least Squares   F-statistic:                     109.3
Date:                Wed, 14 Nov 2018   Prob (F-statistic):           5.72e-70
Time:                        11:43:47   Log-Likelihood:                -3090.9
No. Observations:                 366   AIC:                             6194.
Df Residuals:                     360   BIC:                             6217.
Df Model:                           5
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept   4585.2718    414.985     11.049      0.000    3769.172    5401.372
temp        6666.6567    357.707     18.637      0.000    5963.199    7370.115
hum        -1399.3406    601.266     -2.327      0.021   -2581.776    -216.905
windspeed  -4102.5802    815.500     -5.031      0.000   -5706.323   -2498.837
weathersit  -662.0397    151.428     -4.372      0.000    -959.833    -364.246
workingday   285.0826    127.870      2.229      0.026      33.616     536.549
==============================================================================
Omnibus:                        4.108   Durbin-Watson:                   0.761
Prob(Omnibus):                  0.128   Jarque-Bera (JB):                3.865
Skew:                          -0.243   Prob(JB):                        0.145
Kurtosis:                       3.133   Cond. No.                         31.3
==============================================================================
```

At the 5% significance level, all variables are significant. This means that the analytical methods used in EDA are considered accurate and effective.

Model:

cnt=4585.2718+6666.6567×temp -1399.3406×hum-4102.5802×windspeed

-662.0397×weathersit+285.0826×workingday

Temperature has positive effect with count of rental sharing-bike, while humidity and wind speed have negative effect.

Working day or not and weather situation are dummy variables. Working day and good weather situation have a positive effect with counts whereas non-working day and worse weather situation have a negative effect.