作者：神秘的豆包  
链接：https://zhuanlan.zhihu.com/p/38046021  
来源：知乎  
著作权归作者所有。商业转载请联系作者获得授权，非商业转载请注明出处。

* 从时间维度分析
  + 1.一天中的哪个时间段是用车高峰?
  + 2.一周7天中的用车情况。
  + 3.不同月份的用车情况对比。
  + 4.2011与2012年用车情况分析。
* 从用户维度分析
  + 1.会员与非会员的单车需求量对比。
  + 2.会员与非会员的用车时间是否有差异。
  + 3.会员与非会员在一周中的每一天用车量。
* 从季节维度分析
  + 用车频次是否随着季节的变化而改变?

**1.加载包和读取数据**

import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

import seaborn as sns

%matplotlib inline

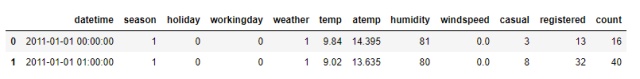
plt.rcParams['font.sans-serif'] = ['SimHei']

plt.rcParams['axes.unicode\_minus'] = False

plt.style.use('ggplot')

df = pd.read\_csv('train.csv')

df.head(2)

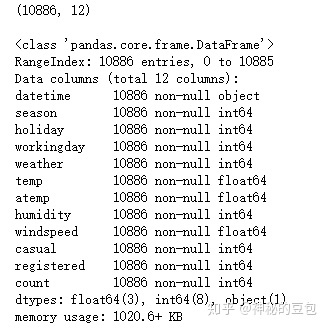


* datatime : 日期时间
* season （季节）: 1-春天；2-夏天；3-秋天；4-冬天
* holiday （是否为节假日）
* workingday （工作日）: 1为工作日 ；0为周末
* weather（天气）： 1-晴天,多云；2-雾天,阴；3-小雪,小雨； 4-大雨,大雪,大雾
* temp : 气温摄氏度
* atemp : 体感温度
* humidity : 湿度
* windspeed : 风速
* casual : 非注册用户个数
* registered : 注册用户个数
* count : 给定日期时间（每小时）总租车人数

# 查看数据类型，各个字段是否有缺失值

print(df.shape,'\n')

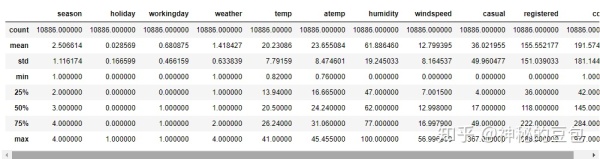
df.info()



数据集一共有10886条数据，并无缺失值。

# 描述性统计

df.describe()



观察数据，从 count 字段看，平均每小时有191人用车，最多的时候每小时有977个用户用车，属于用车高峰期。

结合分位数和平均值看，平均值大于中位数，说明数据是右偏得，也就是肯定存在少部分时刻的用车高峰期，用车人数较多。

casual(非会员),registered(会员)字段同理。

**2.数据预处理**

**2.1 数据变换**

因为数据中并不含有缺失值，我只需要将datetime字段中的日期与时间拆成两列，方便后面的分析。

同时把其更改为日期格式。

df['date'] = df['datetime'].apply(lambda x:x.split(' ')[0])

# 获取小时数

df["hour"] = df['datetime'].apply(lambda x: x.split()[1].split(":")[0])

del df['datetime']

# 获取星期数

df['day'] = df['date'].apply(lambda x:pd.to\_datetime(x,format='%Y-%m-%d')).dt.weekday

# 获取月份数

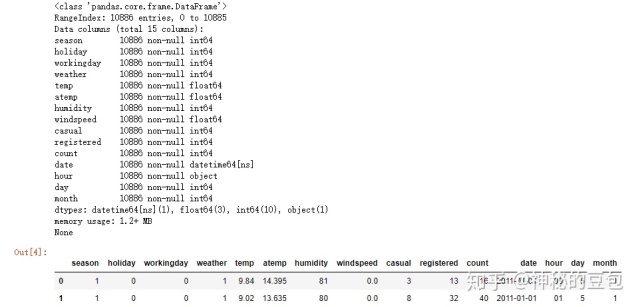
df['month'] = df['date'].apply(lambda x:pd.to\_datetime(x,format='%Y-%m-%d')).dt.month

# 将date字段改为时间格式

df['date'] = pd.to\_datetime(df['date'],format = '%Y-%m-%d')

print(df.info())

df.head(2)



**3.探索性数据分析**

**3.1从时间维度分析**

* **3.1.1 一天中的哪个时间段是用车高峰?**

对于**工作日**和**非工作日**，我认为车辆的需求量是不同。

所以进一步分析工作日和非工作日的用车情况。

# 工作日用车情况

plt.figure(figsize=(15,6))

df[df['workingday']==1].groupby('hour').mean()['count'].plot(linestyle = '-',

linewidth = 2,

color = 'g')

# 非工作日用车情况

df[df['workingday']==0].groupby('hour').mean()['count'].plot(linestyle = '-',

linewidth = 2,

color = 'y')

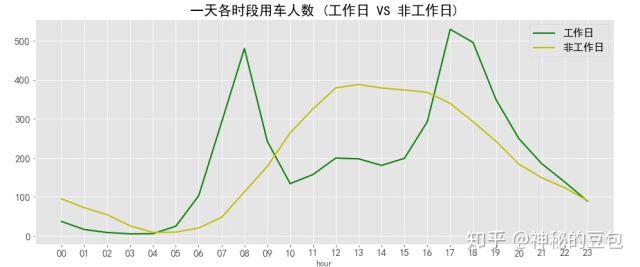
plt.xticks(range(24),df['hour'],fontsize=15)

plt.yticks(fontsize=15)

plt.legend(['工作日','非工作日'],fontsize=15)

plt.title(u'一天各时段用车人数 (工作日 VS 非工作日)',fontsize=20)

plt.show()



对比观察：

1.在工作日，用车量在早8点和傍晚17点是两个用车高峰，正是上下班时期，符合常理；而在中午12点左右也有一小段上升，推测可能是外出吃饭的人。

2.在非工作日，折线图有点近似正态分布，没有上下班时期的用车高峰情况，而峰值集中在了13点左右，说明12-16点时的单车需求量比较大。

3.由此也可以推测，共享单车的**目标客户**应该是普通的上班族居多。

* **3.1.2 一周7天中的用车情况**

plt.figure(figsize=(10,6))

plt.bar(range(7),df.groupby('day').mean()['count'],width=0.3)

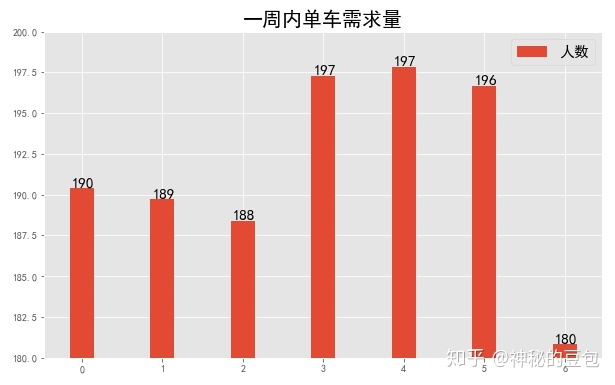
plt.ylim([180,200])

plt.title('一周内单车需求量',fontsize=20)

plt.legend(['人数'],fontsize=15)

for x,y in enumerate(df.groupby('day').mean()['count']):

plt.text(x-0.12,y+0.3,'%i'%y,va='center',fontsize=15)



观察得知，一周中用车量最低的时间是周六，需求量最大的时期是周四。但是仔细观察纵轴，极差只有17，说明一周7天的用车数量还是比较稳定的。

* **3.1.3 不同月份的用车情况对比**

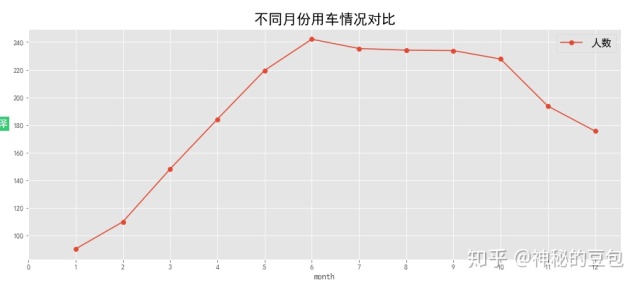
plt.figure(figsize=(15,6))

df.groupby('month').mean()['count'].plot(marker = 'o')

plt.xticks(range(0,13))

plt.title('不同月份用车情况对比',fontsize=20)

plt.legend(['人数'],fontsize=15)



从图上得知，单车需求量在1月份持续上升，在6-10月份迎来高峰，10月之后持续下降。

但是令人不解的是1-3月份的单车需求量并不是很大，1-3月正是春季，然而单车的使用量却比冬天的需求还要低，我猜测是不是有11年与12年的数据收集出了问题，带着这个疑问，我又分别做出了两年的折线图。

* **3.1.4 2011与2012年用车情况分析**

plt.figure(figsize=(15,6))

# 2011年

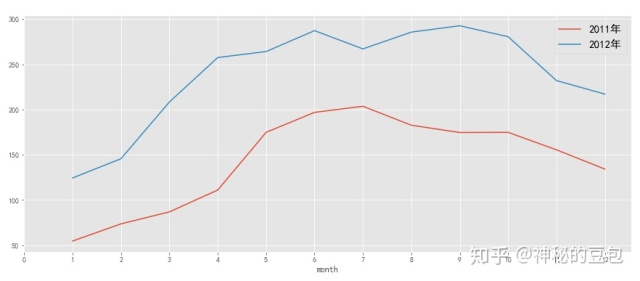
df[df['date'].map(lambda x:x.year)==2011].groupby('month')['count'].mean().plot()

# 2012年

df[df['date'].map(lambda x:x.year)==2012].groupby('month')['count'].mean().plot()

plt.xticks(range(0,13))

plt.legend(['2011年','2012年'],fontsize=15)



1. 经过对比发现，两年单独的折线图与总体的趋势图走向一致，应该不是数据出现了问题，如果想要进一步分析，还需要深入研究当前城市情况，在此就不做讨论了。
2. 虽然两年的总体趋势相似，但2012年的单车使用量明显高于2011年，说明11年的市场运营的效果非常好，用户增长明显。

**3.2 从用户维度分析**

**3.2.1 会员与非会员的单车需求量对比**

df\_registered = df['registered'].sum() / df['count'].sum()

df\_casual = df['casual'].sum() / df['count'].sum()

plt.figure(figsize=(15,6))

plt.axis('equal')

plt.pie(x = [df\_registered,df\_casual],

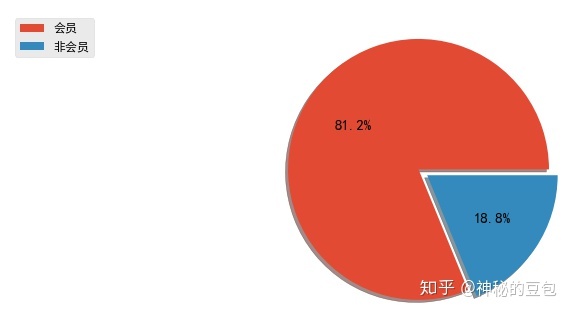
autopct = '%.1f%%',

textprops = {'fontsize' : 15},

explode = [0.08,0],

shadow=True)

plt.legend(['会员','非会员'],fontsize=12,loc = 'upper left')



能够看出，会员还是我们共享单车的使用主体的，占8成左右。

我们再看看会员非会员在一天中的出行时间是否相同。

**3.2.2 会员与非会员在24小时内用车时间是否有差异**

plt.figure(figsize=(15,6))

df.groupby('hour')['registered'].mean().plot()

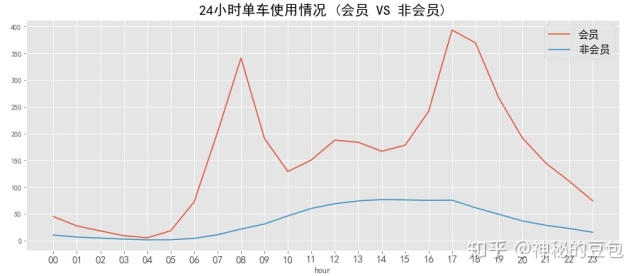
df.groupby('hour')['casual'].mean().plot()

plt.xticks(range(24),df['hour'],fontsize=15)

plt.legend(['会员','非会员'],fontsize = 15)

plt.title('24小时单车使用情况 (会员 VS 非会员)',fontsize=20)

plt.show()



可以看到，非会员的折线图中并没有出现像会员那样的两个峰值，但用车量在上下班也在增加，说明在非注册用户中，也有很大一部分的上班族。我认为对于运营来说，可以进一步加大推广力度，将非会员转化为会员，

**3.2.3 会员与非会员在一周中的用车情况**

df\_week\_registered = df.groupby('day').mean()['registered']

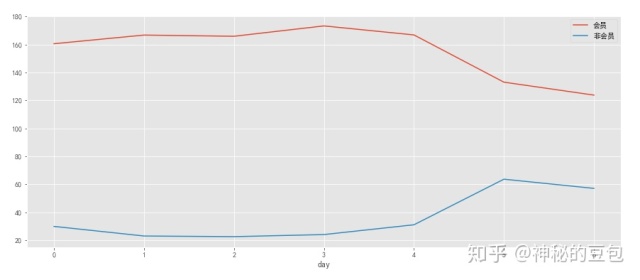
df\_week\_casuual = df.groupby('day').mean()['casual']

plt.figure(figsize=(15,6))

df\_week\_registered.plot()

df\_week\_casuual.plot()

plt.legend(['会员','非会员'])



会员与非会员在一周中的用车频率并不相同，会员在周五周六用车人数最低，而非会员恰恰相反，在周五周六人数较多。

**3.3 从季节维度分析**

* **用车频次是否随着季节的变化而改变?**

plt.figure(figsize=(15,6))

df[df['season']==1].groupby('hour').mean()['count'].plot()

df[df['season']==2].groupby('hour').mean()['count'].plot()

df[df['season']==3].groupby('hour').mean()['count'].plot()

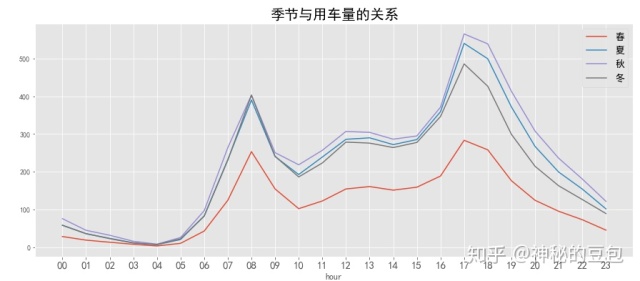
df[df['season']==4].groupby('hour').mean()['count'].plot()

plt.legend(['春','夏','秋','冬'],fontsize=14)

plt.xticks(range(24),df['hour'],fontsize=15)

plt.title('季节与用车量的关系',fontsize=20)

plt.show()



观察发现：

1. 不论季节，每天早晨的7-9点、傍晚16点到19点都为租车的高峰期。
2. 白天9点-16点，可能是因为工作时间的原因，用车人数相对高峰期下降。
3. 春季用车人数最少，秋季用车人数最多。