**Import tools set**

import numpy as np

import pandas as pd

import seaborn as sns

import matplotlib.pyplot as plt

%matplotlib inline

import warnings

warnings.filterwarnings('ignore')

sns.set\_style("whitegrid")

**Import data**

link = <https://ml101-khanhnguyen.s3-ap-southeast-1.amazonaws.com/devc/Online_Retail.csv>

latin1 = pd.read\_csv(link,encoding = 'unicode\_escape')

latin1[:5]

latin1.info()

**Check for NaN values**

latin1.isna().sum()

**Examine few examples of NaN values**

latin1[latin1['Description'].isnull()]

**Exclude negative Quatity entries**

temp=latin1[latin1['Quantity']<0].index.values

latin1.drop(temp,axis=0,inplace=True)

**Exclude negative Price entries**

temp=latin1[latin1['UnitPrice']<0].index.values

latin1.drop(temp,axis=0,inplace=True)

**The customer with the highest number of orders comes from the United Kingdom (UK)**

latin1[latin1['Country']=='United Kingdom']['CustomerID'].value\_counts().head().plot(kind='barh')

**The customer with the highest money spent on purchases comes from Netherlands**

Spent = latin1['Quantity'] \* latin1['UnitPrice']

latin1.insert(loc=8,column='Spent',value=Spent)

latin2 = latin1[latin1['Country']=='Netherlands']

latin3 = latin2.groupby(['CustomerID']).sum()

latin3.head().plot(kind='barh')

**On which year had the highest sales?**

latin3 = latin1[['Quantity','Spent','InvoiceDate','Country']].copy()

latin3['InvoiceDate'] = pd.to\_datetime(latin3['InvoiceDate'], errors='coerce')

latin3['Month'] = latin3['InvoiceDate'].dt.month

latin3['Year'] = latin3['InvoiceDate'].dt.year

**How many orders (per hour)?**

year = latin3.groupby('Year').sum()

sales = year.sort\_values('Spent',ascending=False)

sales = sales['Spent']

sales

latin3['Date'] = latin3['InvoiceDate'].dt.date

latin3['Hour'] = latin3['InvoiceDate'].dt.hour

date = latin3.groupby('Date').count()

ma = latin3['Hour'].max()

mi = latin3['Hour'].min()

per\_hour = date['InvoiceNo'] // (ma-mi)

per\_hour

**Make a plot about number of orders per hour**

sns.distplot(per\_hour,bins=20)

**How many orders (per month)?**

month = latin3.groupby('Month').count()

per\_month = month['InvoiceNo'] // 12

per\_month

**Make a plot about number of orders per month**

Mo=[1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12]

plt.bar(Mo,per\_month)

**Top 10 items most sales**

pro\_spent = latin1.groupby(['StockCode']).sum()

pro\_spent.sort\_values('Spent',ascending=False,inplace=True)

print (pro\_spent['Spent'].head(10))

**Create a histogram with the 10 countries that have the most 'Quantity' ordered except UK**

latin4 = latin1[latin1['Country']!='United Kingdom']

con\_quantity=latin4.groupby(['Country']).sum()

con\_quantity.sort\_values('Quantity',ascending=False,inplace=True)

cq = con\_quantity['Quantity'].head(10)

sns.distplot(cq,bins=8)