The Report of Homework 1.1

Individual Contribution			
CWID	Name	Contribution (description)	Percent Contribution
A20563460	Yuxuan Yang	Write code	50.0%
A20563458	Chang Li	Write a report	50.0%

• Describe the dataset

Our group made two sets of data, one of which was the insurance cost and BMI and smoking status data set.
The variables are BMI (body mass index) and Charges (insurance cost), and the population is also divided into smokers and non-smokers, respectively. There are eight sets of data, namely:

```
'bmi': [25, 30, 22, 28, 35, 24, 31, 27],
'charges': [1000, 1500, 800, 1200, 2000, 900, 1600, 1300],
'smoker': ['yes', 'no', 'no', 'yes', 'yes', 'no', 'yes', 'no']
```

The second data is a data set of daily global playback volume of popular songs.

This dataset records the daily global play of a popular song between 2017 and 2018, with variable dates and views.

```
'date': ['2017-01-01', '2017-03-01', '2017-05-01', '2017-07-01',
'2017-09-01', '2017-11-01', '2018-01-01'],
'value': [10, 20, 15, 30, 10, 20, 35]
```

• visualization(s)

(a)Linear diagram

Title: BMI vs Insurance Changes By Smoking Status
BMI: as an independent variable, it is located on the X-axis.
The lowest to the highest ones are arranged in ascending order.
Insurance cost: on the Y axis. In monetary units.
• Smoking status: As a categorical variable, use different colors to distinguish between smokers and non-smokers Details: Color distinguishes the approximate range of the data

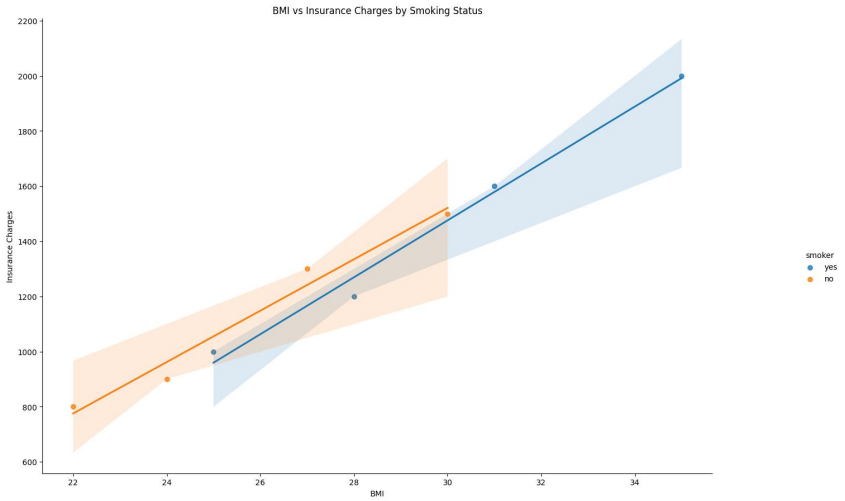

Figure 1

(b) line chart

title: Daily Global Streams of Prpular Songs (2017-2018)

• Date: As an independent variable, it is located on the X-axis. Data are arranged from the earliest to the latest.
• Play volume: as a dependent variable, it is located on the Y-axis. In the number of times.

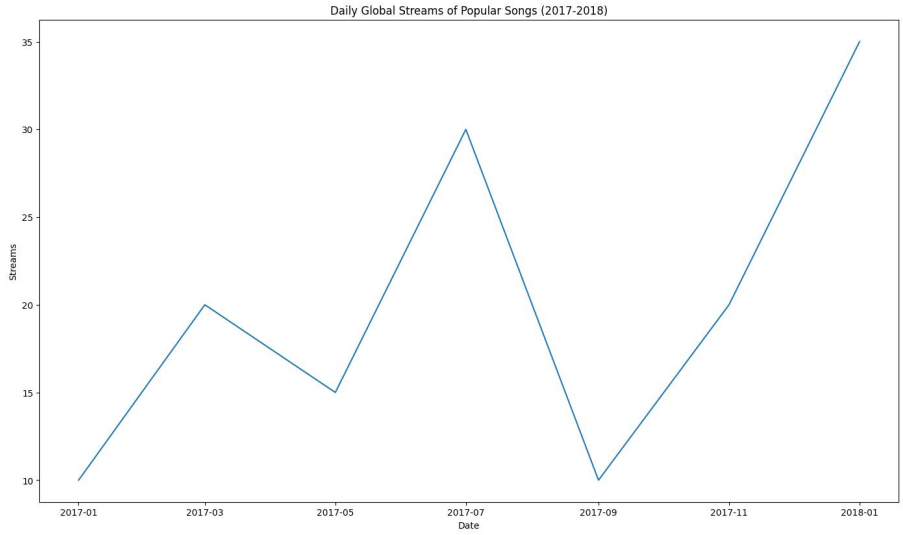

Figure 2

·Method and library used to create the visualization

1、Import necessary libraries

Using import pandas as pd and import seaborn as sns to import the libraries required for data processing and visualization.

2、Select the visualization type

（a）Linear diagram

Suitable for showing the relationship between two continuous variables and can be colored according to the third categorical variable. For example, when analyzing the relationship between BMI and insurance costs, we colored it according to smoking status.

```
sns.lmplot(x="bmi", y="charges", hue="smoker", data=insurance_data)
```

(b)line chart

For trends used to show changes in the time series data.

```
sns.lineplot(x='date', y='value', data=spotify_data)
```

3、Add the title and label & display the chart

·conclusion

It can be reasonably inferred from Figure 1: for both smokers and non-smokers, the insurance cost is also increasing with the increase of BMI. This suggests that insurers consider customers' weight in pricing, believing that customers with higher weight may have a higher health risk and therefore pay higher insurance costs. Furthermore, insurance costs were generally higher among smokers than among non-smokers, further emphasizing the impact of a healthy lifestyle on insurance costs.

In Figure 2, we found that the number of song plays was significantly increased or decreased on certain dates (July of 17 to January of the next year), and was relatively stable in some periods (March-May of 17). It is obvious that the daily global broadcast volume increases during holidays (winter and summer vacations), which shows that new songs can be released during these periods, so that there is more traffic.