

นายภากรณ์ ธนประชาพันธ์ 62010694

Homework #2

สถิติความล่าช้าของเที่ยวบินภายในประเทศสหรัฐอเมริกาเดือนสิงหาคม

2018 August 2018 Nationwide Airplane Delay Statistic

เลือกใช้ Column : DEP_DELAY (Departure Delay (HHMM)) และ ARR_DELAY (Arrival Delay (HHMM))

เพื่อหาความสัมพันธ์ว่า และเวลาล่าช้าขาออกมีผลต่อเวลาล่าช้าขาเข้าหรือไม่

ข้อมูลสถิติต่างๆที่ได้จากชุดข้อมูลนี้

August 2018 Nationwide Airplane Delay Statistic					

	DEP_DELAY		ARR_DELAY		
count	23140.000000		23140.000000		
mean	17.939283		17.132930		
std	52.561541		55.989343		
min	-36.000000		-60.000000		
25%	-5.000000		-11.000000		
50%	-1.000000		-1.000000		
75%	17.000000		22.000000		
max	1032.000000		1063.000000		

Mode					

	FL_DATE	ORIGIN	DEST	DEP_DELAY	ARR_DELAY
0	8/1/2018	ATL	ATL	-5	-9

Measures of spread (Dispersion)					

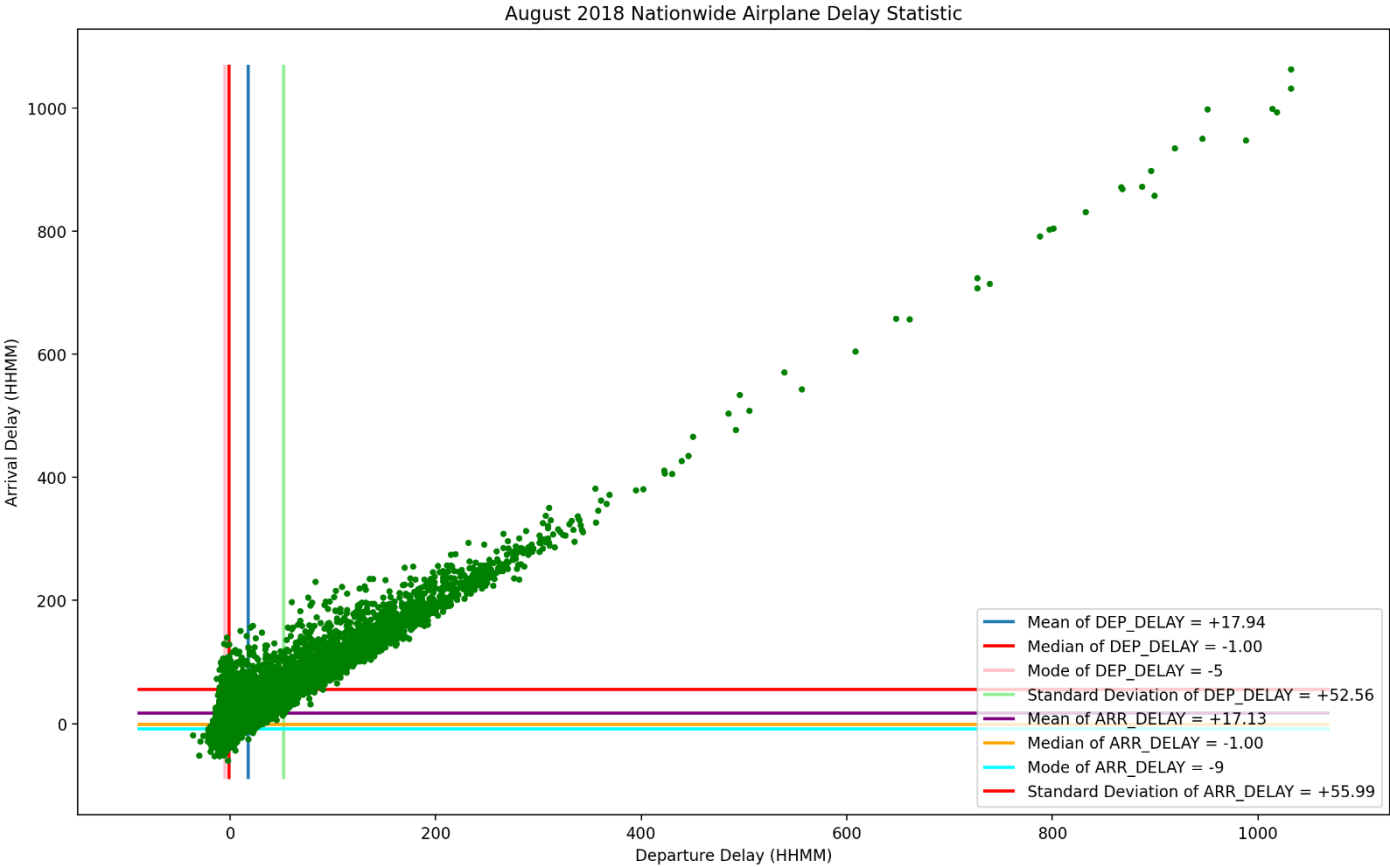
1068 (HHMM)					

Simple Variance					

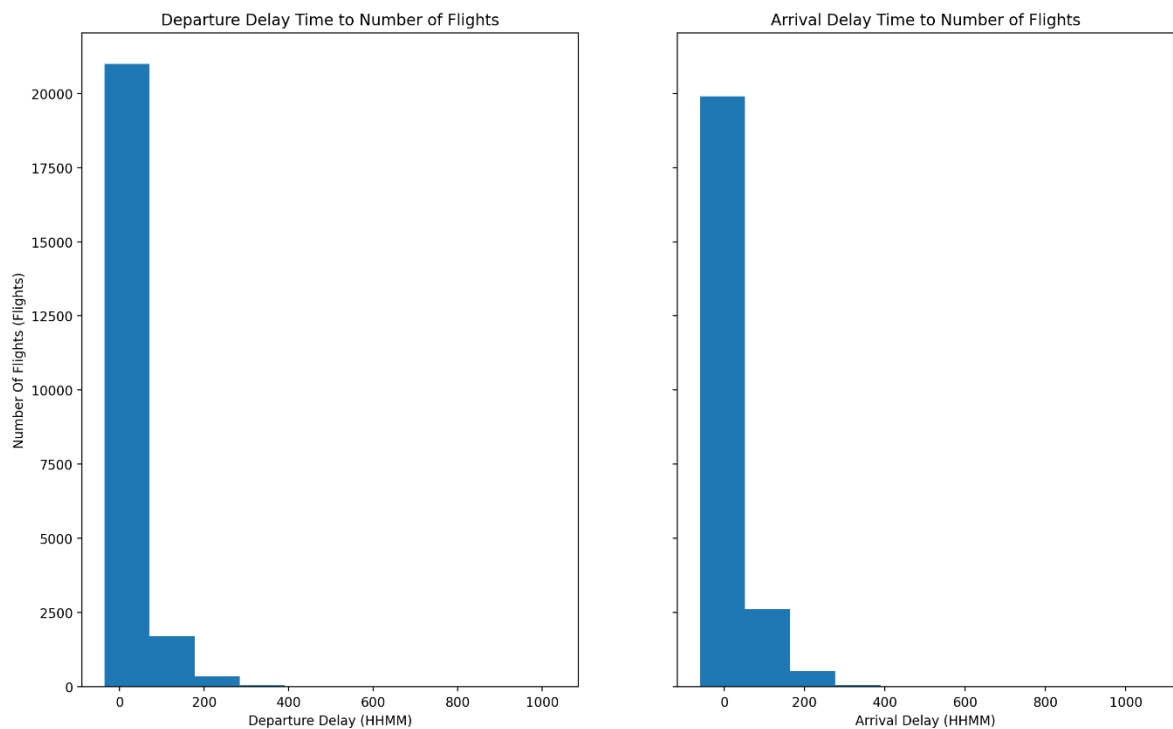
DEP_DELAY : 2762.715575093635					
ARR_DELAY : 3134.8065649918576					

โดยค่าสถิติพื้นฐานอื่นๆสามารถดูได้จากกราฟ Scatter Plot ด้านล่าง

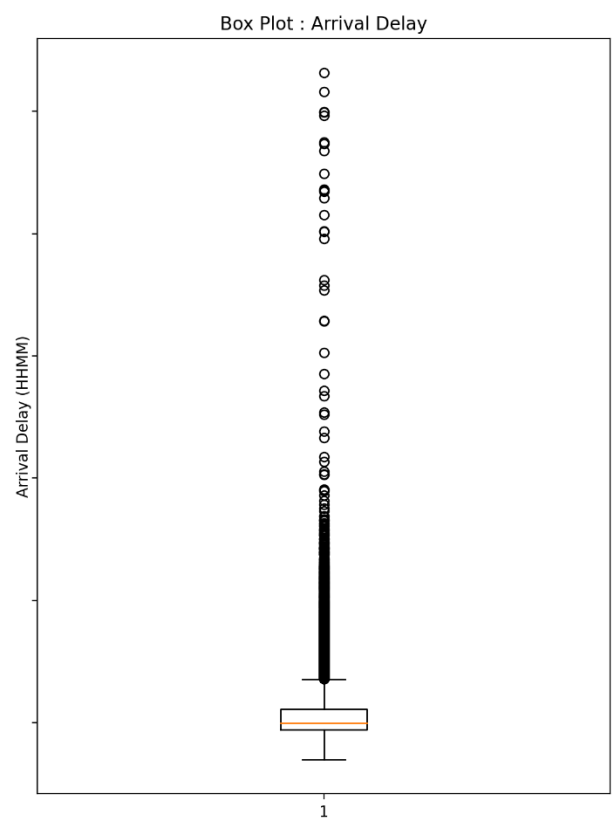
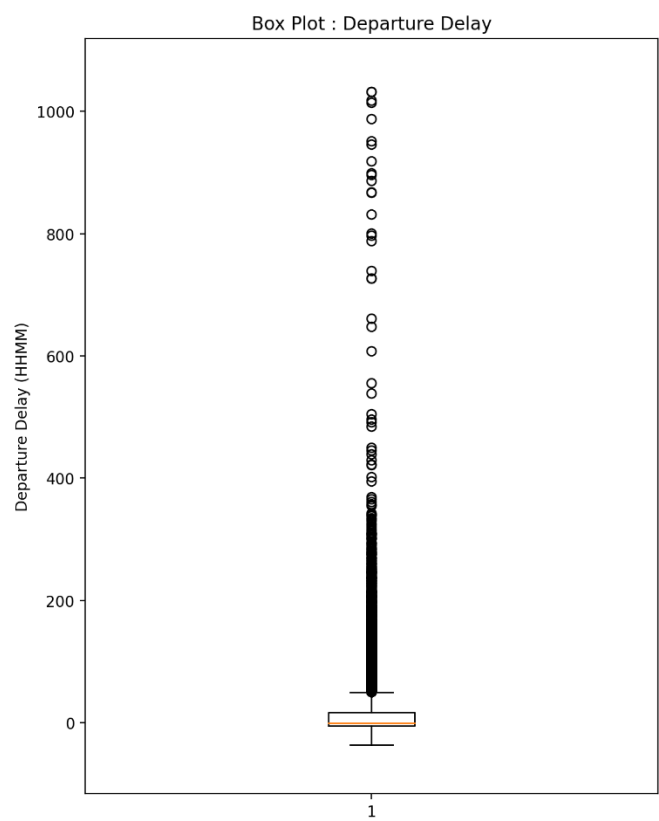
Scatter Plot



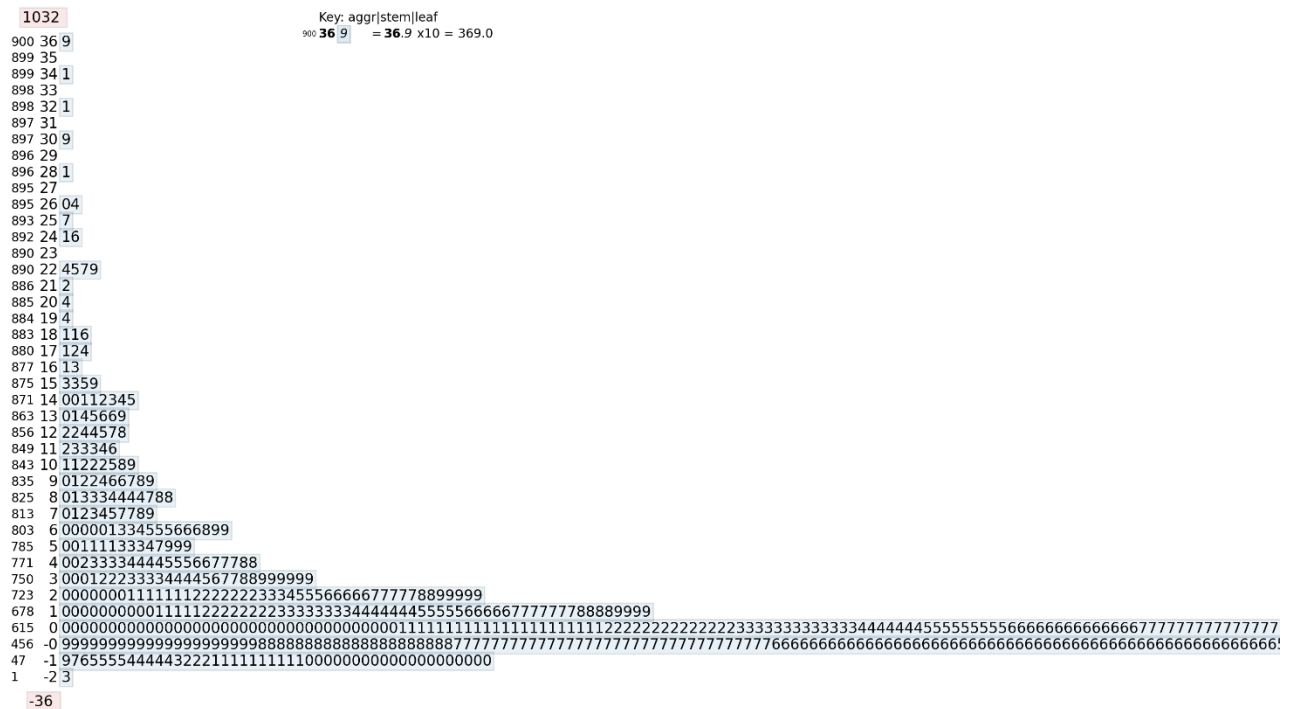
Histogram



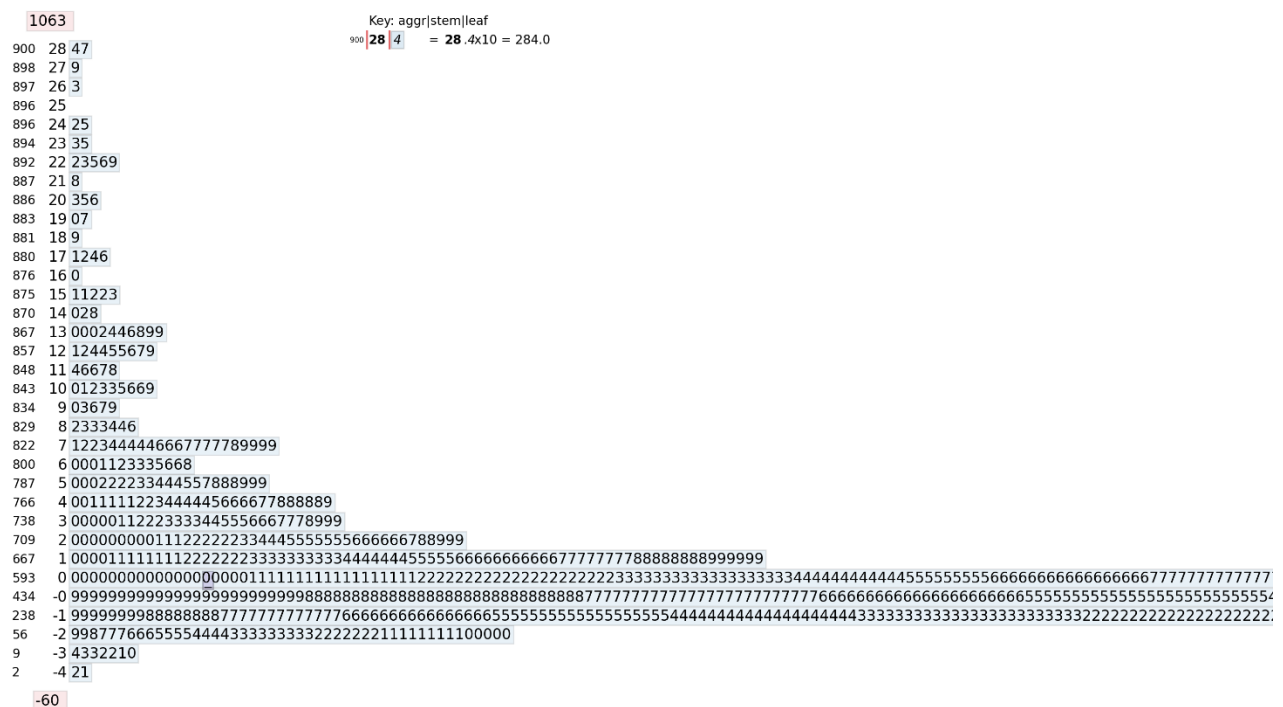
Box Plot



Stem and Leaf: Departure Delay (HHMM)



Stem and Leave: Arrival Delay (HHMM)



ให้ตัวแปรต้นเป็น : DEP_DELAY (Departure Delay (HHMM)) เวลาล่าช้าขาออก

ตัวแปรตามเป็น : ARR_DELAY (Arrival Delay (HHMM)) เวลาล่าช้าขาเข้า

เหตุผล เพราะต้องการที่จะศึกษาประสิทธิภาพของการเดินทาง ด้วยเครื่องบินพาณิชย์ที่เวียนภายในของประเทศ สหรัฐอเมริกาว่าเวลาล่าช้าขาออกมีผลต่อเวลาล่าช้าขาเข้าหรือไม่ หรือว่าสองอย่างนี้ไม่มีความเกี่ยวข้องโดยกันสิ้นเชิง มีความเป็นไปได้หรือไม่ที่เครื่องบินออกช้าจะถึงจุดหมายตรงเวลา หรือเร็วกว่าเวลากำหนดเดิม

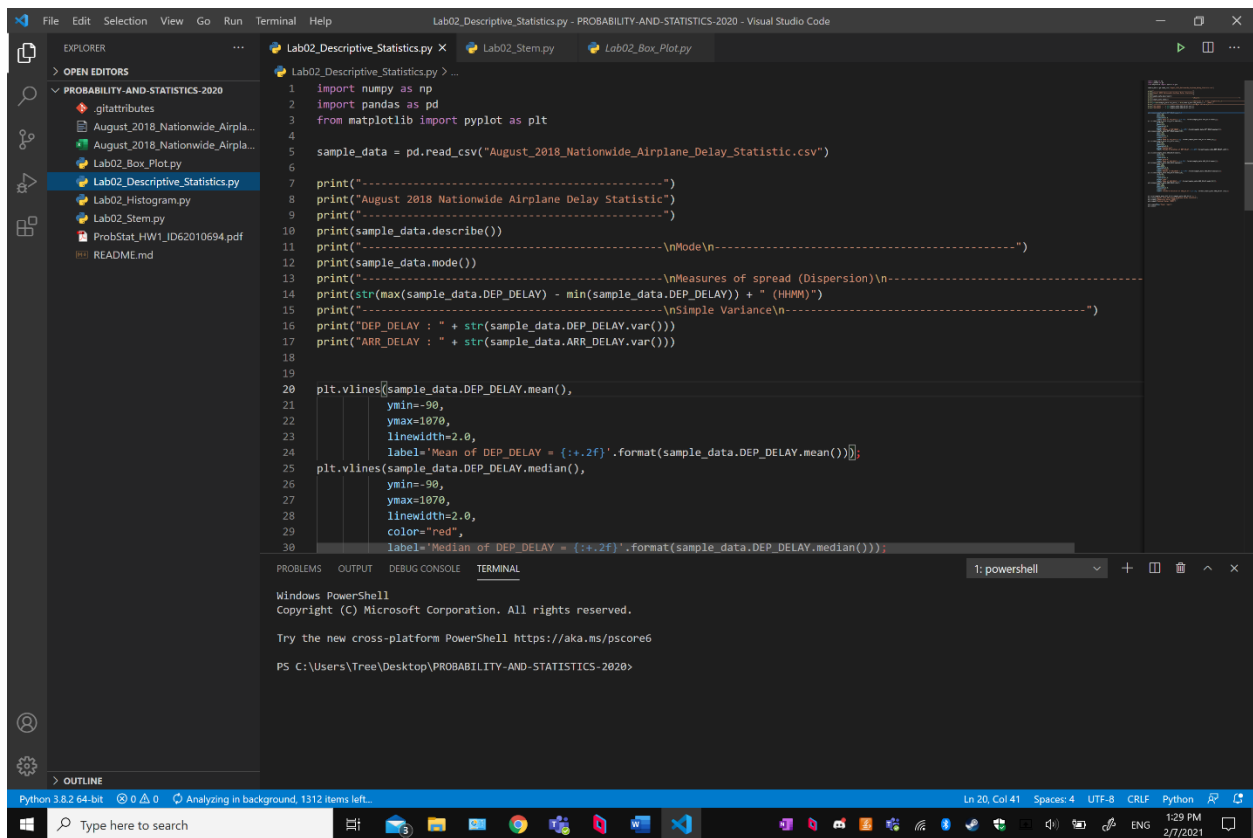
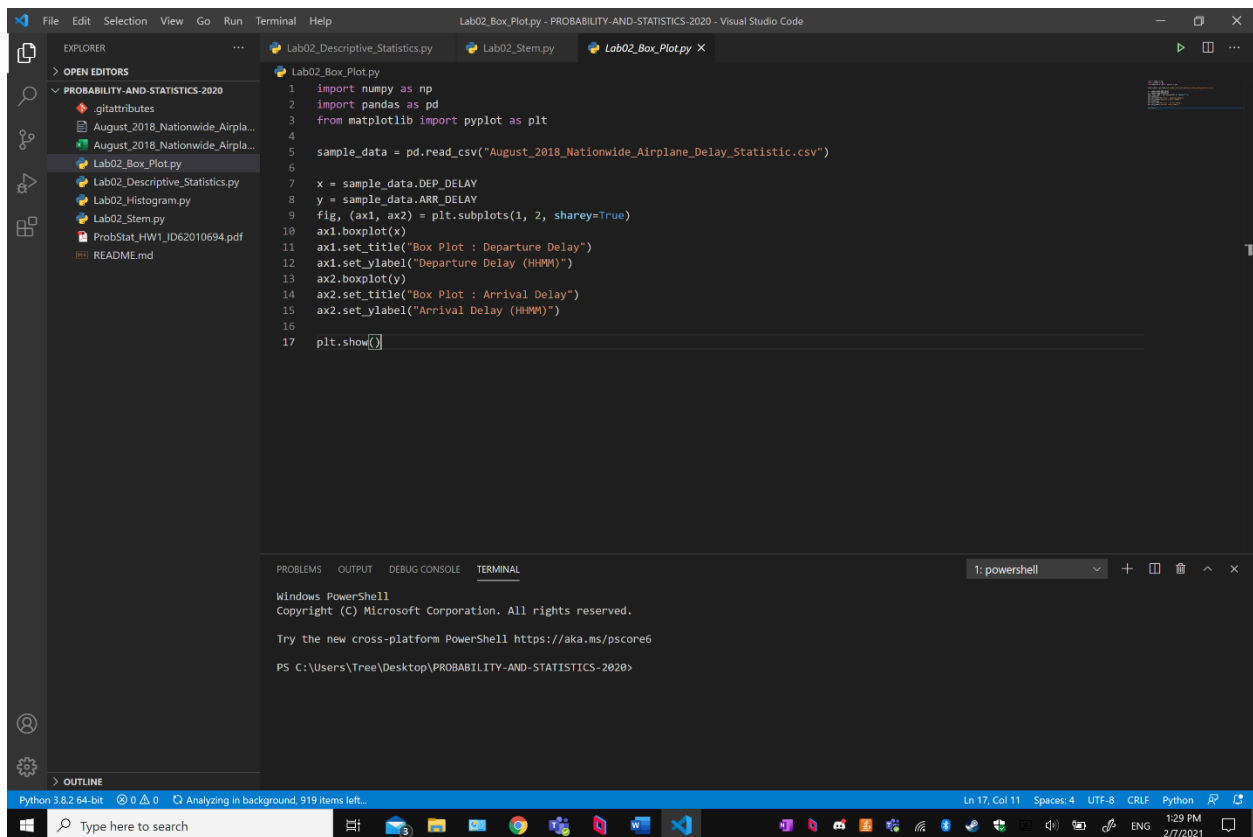
Outlier: เนื่องจากชุดข้อมูลนี้มี Outlier จำนวนมากจึงจะขอยกตัวอย่างค่าที่เป็น Outlier สูงสุดแทน

ตามแนวแกน X (ค่าเวลาล่าช้าขาออก) : 1032 (ล่าช้าไป 10 ชั่วโมง 32 นาที)

ตามแนวแกน Y (ค่าเวลาล่าช้าขาเข้า) : 1063 (ล่าช้าไป 11 ชั่วโมง 3 นาที) *คาดว่าเป็นการผิดพลาดของการเก็บข้อมูล

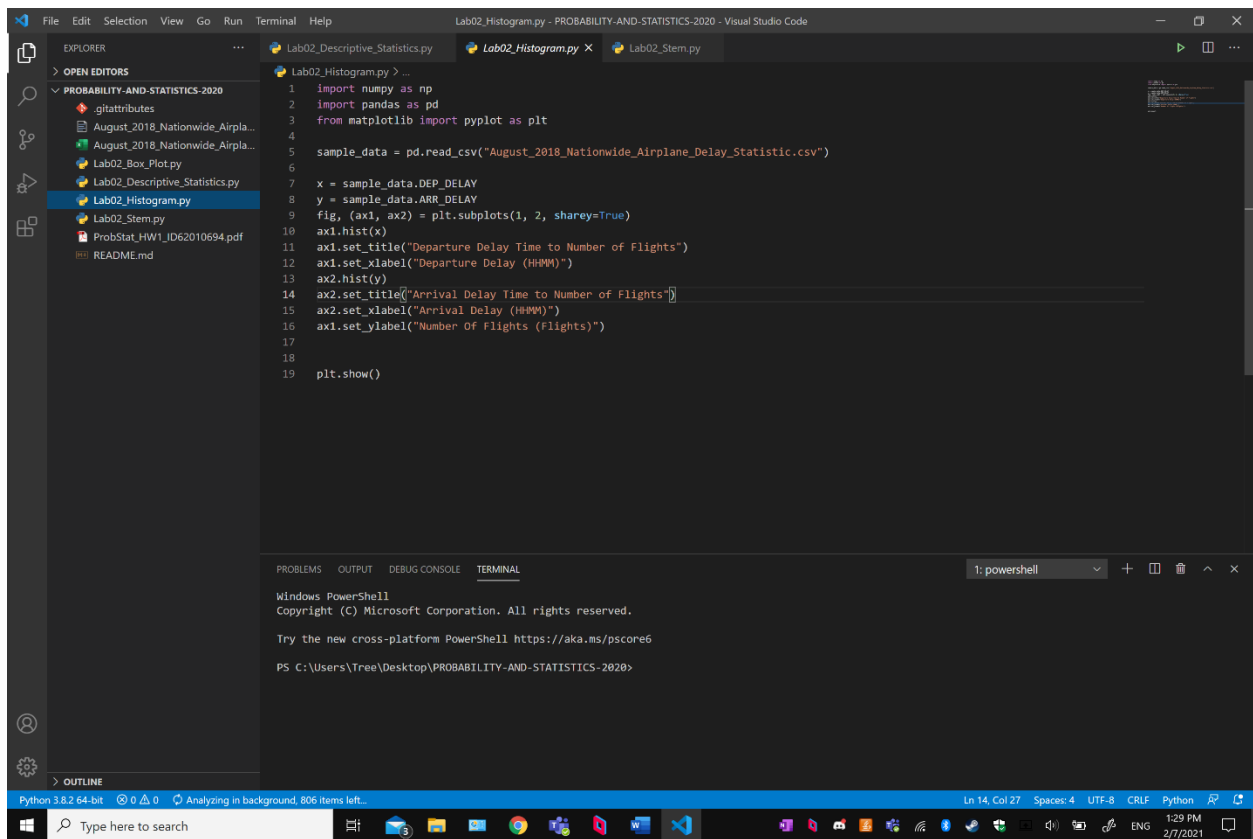
บทวิเคราะห์ข้อมูลจากกราฟ : จากชุดข้อมูลที่ได้นี้มาจะสามารถเห็นได้ว่า ค่าเวลาล่าช้าขาออก (Departure Delay (HHMM)) ส่งผลโดยตรงกับ ค่าเวลาล่าช้าขาเข้า (Arrival Delay (HHMM)) อย่างชัดเจนในรูปแบบแปรผันตรง หากเที่ยวบินออกช้า เวลาถึงจะช้าตามไปด้วย มีความเป็นไปได้ที่เครื่องบินออกช้าจะถึงจุดหมายตรงเวลา หรือเร็วกว่าเวลากำหนดเดิม แต่จะไม่มีทางเร็วไปมากกว่าสองชั่วโมงอย่างแน่นอน มีเที่ยวบินที่ล่าช้าถึง 10 ชั่วโมงอยู่จริง และจากการหาค่า Mode เราสามารถระบุสนามบินทั้งต้นทางและปลายทางที่มีประวัติเครื่องบินล่าช้าสุดในเดือนสิงหาคม 2018 ได้ นั่นคือท่าอากาศยานนานาชาติฮาร์ตสฟิลด์-แจ็คสัน แอตแลนตา (Hartsfield-Jackson Atlanta International Airport) ซึ่งมี IATA Airport Code คือ ATL และจากทั้งชุดข้อมูลนี้ ค่าเวลาล่าช้าขาออก (Departure Delay (HHMM)) และ ค่าเวลาล่าช้าขาเข้า (Arrival Delay (HHMM)) ที่พบได้มากที่สุดคือ -5 (ออกไวไป 5 นาที) และ -9 (ถึงไวไป 9 นาที) ตามลำดับ สุดท้ายจากชุดข้อมูลนี้เราสามารถบอกได้ว่า เวลาล่าช้าของเครื่องบินนั้นจะอยู่ในช่วงถึงช้า 2 ชั่วโมงและออกช้า 2 ชั่วโมง เพราะข้อมูลจะเกาะกลุ่มกันอยู่ในช่วงซ้ายล่างของ Scatter Plot เที่ยวบินที่ล่าช้ามากกว่านี้มีอยู่จริง แต่ถือว่าเป็นจำนวนน้อยมากหากเทียบกับจำนวนข้อมูลที่เกาะกลุ่มนี้

เราจึงวิเคราะห์ได้ว่า เวลาล่าช้าทั้งขาออกจะส่งผลกับเวลาล่าช้าขาเข้าในเกือบจะทุกกรณีในรูปแบบแปรผันตรงอย่างเห็นได้ชัด หรือกล่าวอีกนัยหนึ่งคือ หากเครื่องออกช้า จะมีความเป็นไปได้สูงอย่างมากที่จะถึงที่หมายปลายทางล่าช้าอย่างแน่นอน



```
Lab02_Descriptive_Statistics.py > ...
47 ymax=1070,
48 linewidth=2.0,
49 color="red",
50 label="Median of DEP_DELAY = {:.2f}".format(sample_data.DEP_DELAY.median());
51 plt.vlines(sample_data.DEP_DELAY.mode()[0],
52 ymin=-90,
53 ymax=1070,
54 linewidth=2.0,
55 color="pink",
56 label="Mode of DEP_DELAY = {:.2f}".format(sample_data.DEP_DELAY.mode()[0]));
57 plt.vlines(sample_data.DEP_DELAY.std(),
58 ymin=-90,
59 ymax=1070,
60 linewidth=2.0,
61 color="lightgreen",
62 label="Standard Deviation of DEP_DELAY = {:.2f}".format(sample_data.DEP_DELAY.std()));
63
64 plt.hlines(sample_data.ARR_DELAY.mean(),
65 xmin=-90,
66 xmax=1070,
67 linewidth=2.0,
68 color="purple",
69 label="Mean of ARR_DELAY = {:.2f}".format(sample_data.ARR_DELAY.mean()));
70 plt.hlines(sample_data.ARR_DELAY.median(),
71 xmin=-90,
72 xmax=1070,
73 linewidth=2.0,
74 color="orange",
75 label="Median of ARR_DELAY = {:.2f}".format(sample_data.ARR_DELAY.median()));
76 plt.hlines(sample_data.ARR_DELAY.mode()[0],
77 xmin=-90,
78 xmax=1070,
79 linewidth=2.0,
80 color="orange",
81 label="Mode of ARR_DELAY = {:.2f}".format(sample_data.ARR_DELAY.mode()[0]));
82
83 plt.plot(sample_data.DEP_DELAY, sample_data.ARR_DELAY, 'g.')
84 plt.title("August 2018 Nationwide Airplane Delay Statistics")
85 plt.xlabel("Departure Delay (HMM)")
86 plt.ylabel("Arrival Delay (HMM)")
87 plt.legend(loc='lower right')
88 plt.show()
```

```
Lab02_Descriptive_Statistics.py > ...
48 color= purple ,
49 label="Mean of ARR_DELAY = {:.2f}".format(sample_data.ARR_DELAY.mean());
50 plt.hlines(sample_data.ARR_DELAY.median(),
51 xmin=-90,
52 xmax=1070,
53 linewidth=2.0,
54 color="orange",
55 label="Median of ARR_DELAY = {:.2f}".format(sample_data.ARR_DELAY.median()));
56 plt.hlines(sample_data.ARR_DELAY.mode()[0],
57 xmin=-90,
58 xmax=1070,
59 linewidth=2.0,
60 color="cyan",
61 label="Mode of ARR_DELAY = {:.2f}".format(sample_data.ARR_DELAY.mode()[0]));
62 plt.hlines(sample_data.ARR_DELAY.std(),
63 xmin=-90,
64 xmax=1070,
65 linewidth=2.0,
66 color="red",
67 label="Standard Deviation of ARR_DELAY = {:.2f}".format(sample_data.ARR_DELAY.std()));
68
69
70
71 plt.plot(sample_data.DEP_DELAY, sample_data.ARR_DELAY, 'g.')
72 plt.title("August 2018 Nationwide Airplane Delay Statistics")
73 plt.xlabel("Departure Delay (HMM)")
74 plt.ylabel("Arrival Delay (HMM)")
75 plt.legend(loc='lower right')
76 plt.show()
77
```

This screenshot shows the Visual Studio Code interface with the file `Lab02_Histogram.py` open. The Explorer sidebar on the left shows a project named `PROBABILITY-AND-STATISTICS-2020` with various files including `Lab02_Histogram.py`, `Lab02_Descriptive_Statistics.py`, and `Lab02_Stem.py`. The main editor displays the following Python code:

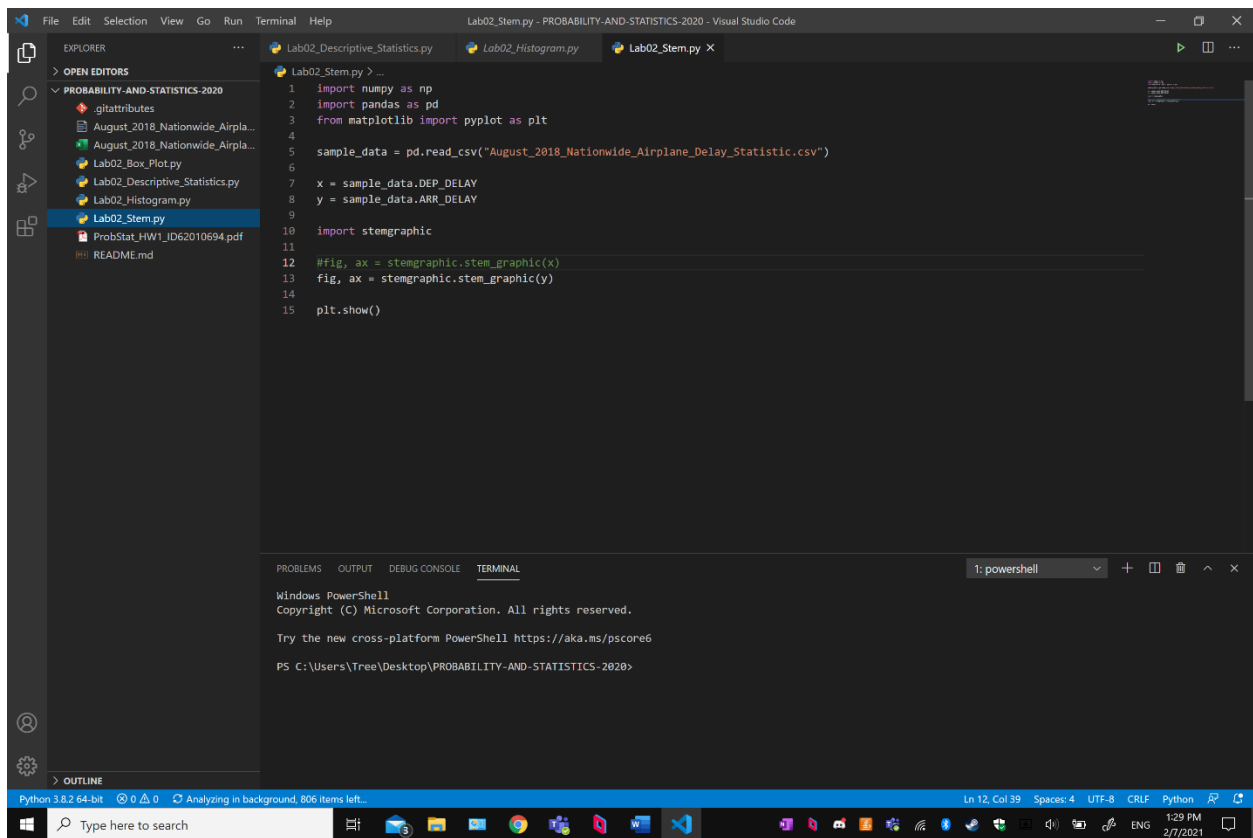
```
1 import numpy as np
2 import pandas as pd
3 from matplotlib import pyplot as plt
4
5 sample_data = pd.read_csv("August_2018_Nationwide_Airplane_Delay_Statistic.csv")
6
7 x = sample_data.DEP_DELAY
8 y = sample_data.ARR_DELAY
9 fig, (ax1, ax2) = plt.subplots(1, 2, sharey=True)
10 ax1.hist(x)
11 ax1.set_title("Departure Delay Time to Number of Flights")
12 ax1.set_xlabel("Departure Delay (H:MM)")
13 ax2.hist(y)
14 ax2.set_title("Arrival Delay Time to Number of Flights")
15 ax2.set_xlabel("Arrival Delay (H:MM)")
16 ax1.set_ylabel("Number Of Flights (Flights)")
17
18
19 plt.show()
```

The bottom of the window shows a terminal window with the following text:

```
Windows PowerShell
Copyright (C) Microsoft Corporation. All rights reserved.

Try the new cross-platform PowerShell https://aka.ms/pscore6

PS C:\Users\Tree\Desktop\PROBABILITY-AND-STATISTICS-2020>
```



This screenshot shows the Visual Studio Code interface with the file `Lab02_Stem.py` open. The Explorer sidebar on the left shows the same project as the first screenshot, but with `Lab02_Stem.py` selected. The main editor displays the following Python code:

```
1 import numpy as np
2 import pandas as pd
3 from matplotlib import pyplot as plt
4
5 sample_data = pd.read_csv("August_2018_Nationwide_Airplane_Delay_Statistic.csv")
6
7 x = sample_data.DEP_DELAY
8 y = sample_data.ARR_DELAY
9
10 import stemgraphic
11
12 #fig, ax = stemgraphic.stem_graphic(x)
13 fig, ax = stemgraphic.stem_graphic(y)
14
15 plt.show()
```

The bottom of the window shows the same terminal window as the first screenshot.