Module 6 Assignment

Rudy Fasano

12 / 07 / 2020

Exercise 1

1.1

importing of required packages as well as airline_costs.csv

In [ ]:
```python
import pandas as pd
import statsmodels.api as sm
import matplotlib as mpl
import matplotlib.pyplot as plt
import numpy as np
from numpy import mean
from numpy import std
from statistics import median
```

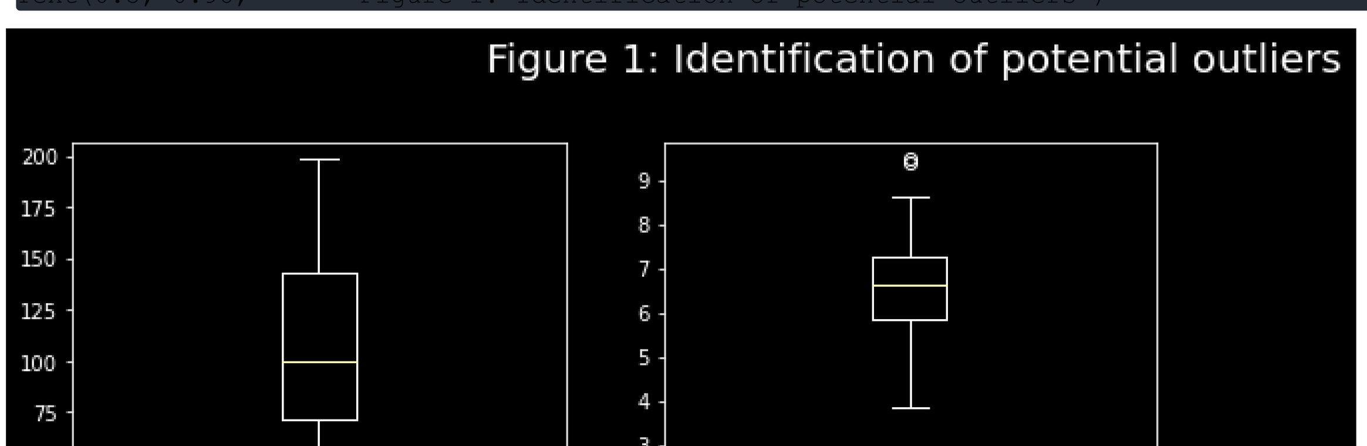
In [ ]:
```python
df = pd.read_csv("/Users/fasan/Desktop/500B/Module6/airline_costs.csv", header=0, sep=',')
df.head()
```

Exploratory analyses were conducted to visualize data as well as to detect null values and potential outliers. No null values were detected, however, several potential outliers were detected in the column of DailyFlightTime by box plot analysis and were replaced by their respective median due to having a small amount of values. The deletion of those rows were believed to be too much of a risk.

In [36]:
```python
df.isnull() ### investigating for null values - none detected

fig = plt.figure(figsize = (15,7)) ### identifying potential outliers with boxplots
plt.style.use('dark_background')
ax1 = fig.add_subplot(231)
ax1.boxplot(x="FlightLength", data=df)
plt.xlabel('Flight Length')
##
ax2 = fig.add_subplot(232) ### various potential outliers detected
ax2.boxplot(x="DailyFlightTime", data=df)
plt.xlabel('Daily Flight Time')
##
plt.suptitle('      Figure 1: Identification of potential outliers',
fontsize=20)
```

Out[36]:
```
Text(0.5, 0.98, '      Figure 1: Identification of potential outliers')
```



In [12]:
```python
df.describe()
```

Out[12]:

	FlightLength	PlaneSpeed	DailyFlightTime	CustomersServed	TotalOperatingCost	Revenue	LoadFactor	AvailableCapacity	TotalAss
count	31.000000	31.000000	31.000000	31.000000	31.000000	31.000000	31.000000	31.000000	31.0000
mean	129.096774	161.258065	6.463226	14491.967742	113.506452	1.732258	0.476290	3.306806	215.3109
std	73.212638	26.851403	1.549586	16824.247836	142.704637	1.191852	0.139171	1.689303	402.6555
min	45.000000	116.000000	2.350000	183.000000	42.300000	0.070000	0.166000	0.422000	2.0300
25%	71.000000	141.500000	5.865000	2500.000000	50.800000	0.800000	0.399000	2.119500	13.2300
50%	100.000000	150.000000	6.600000	6500.000000	75.400000	1.190000	0.505000	2.405000	21.6000
75%	174.500000	181.500000	7.260000	19100.000000	120.750000	2.680000	0.568500	4.598000	167.3100
max	293.000000	216.000000	9.500000	56928.000000	820.900000	4.300000	0.689000	7.544000	1436.5300

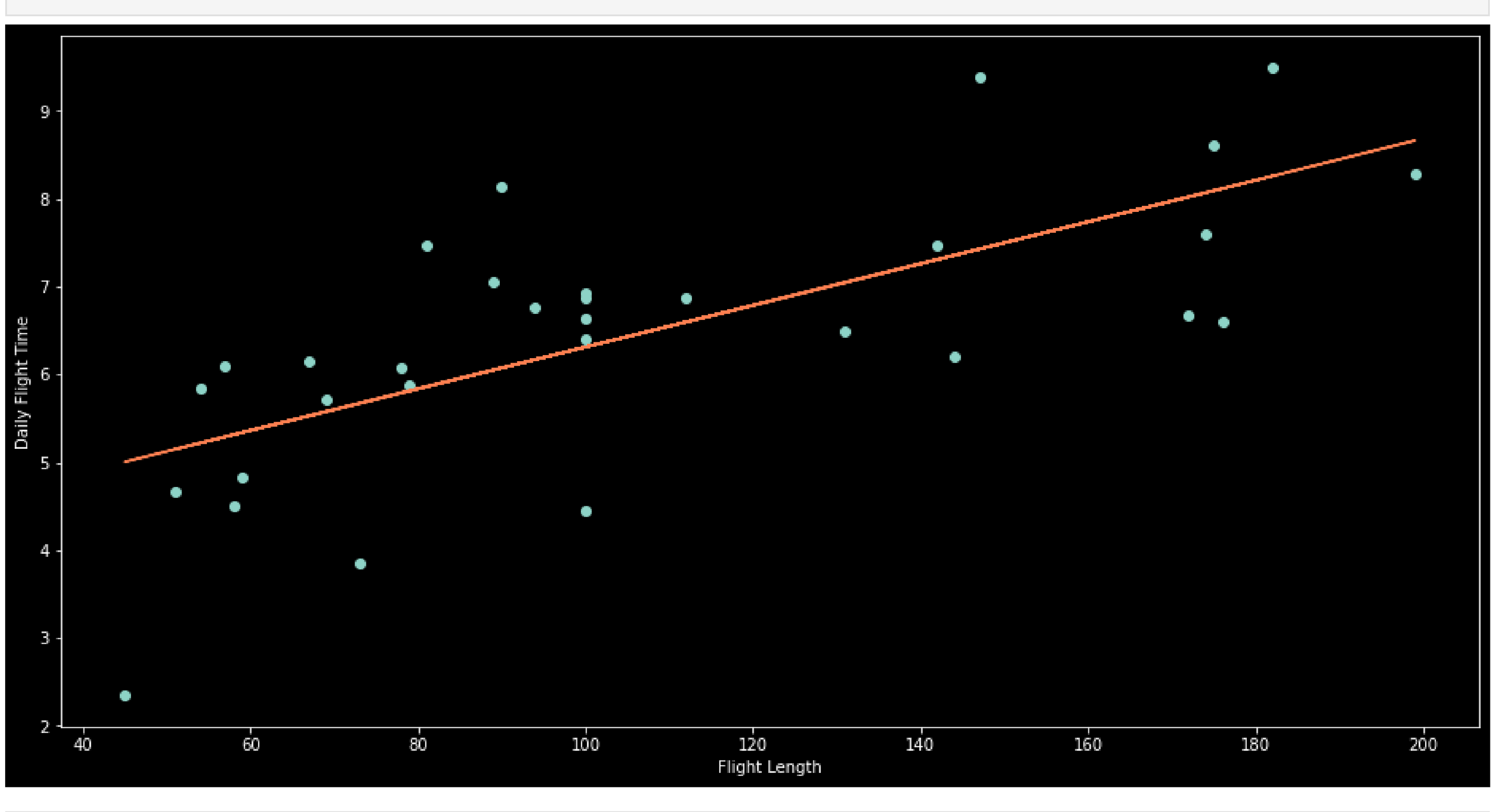
In [30]:
```python
fig, ax = plt.subplots(figsize=(16,8))
ax.scatter(df['FlightLength'], df['DailyFlightTime']) #### further visualization
ax.set_xlabel('Flight Length')
ax.set_ylabel('Daily Flight Time')
ax.set_title('Figure 2: Visualization of data', fontsize=20)
```

Out[30]:
```
Text(0.5, 1.0, 'Figure 1: Visualization of data')
```

In [14]:
```python
df.describe() ## finding the median of flight length = 100.000
##
df['FlightLength'] =df['FlightLength'].mask(df['FlightLength'] > 250, median)
## replacing potential outliers with the median value
```

After the cleaning of data was conducted as well as regression analysis performed, there appears to be a strong positive relationship between flight lengths as well as daily flight times. The daily flight time has been reported to have a coefficient of 0.0552, meaning that as the dependent variable (FlightLength) increased by 1, the predictor variable (DailyFlightTime) will increase by 0.0552. The R-squared is valued to be at 90% (.901) which depicts that approximately 90% of the data is displayed in the model, which is exceptionally high.

In [15]:
```python
fig, ax = plt.subplots(figsize=(16,8))
ax.scatter(df['FlightLength'], df['DailyFlightTime']) #### further visualization
ax.set_xlabel('Flight Length')
ax.set_ylabel('Daily Flight Time')
x=df['FlightLength']
y=df['DailyFlightTime']
m, b = np.polyfit(x, y, 1)
plt.plot(x, m*x+b, c='coral')
plt.show()
```



In [20]:
```python
x = df['FlightLength']
y = df['DailyFlightTime']
mo1 = sm.OLS(y, x).fit()
predictions = mo1.predict(x)
mo1.summary()
```

Out[20]:

OLS Regression Results

Dep. Variable:	DailyFlightTime	**R-squared (uncentered):**	0.920
Model:	OLS	**Adj. R-squared (uncentered):**	0.918
Method:	Least Squares	**F-statistic:**	346.2
Date:	Tue, 08 Dec 2020	**Prob (F-statistic):**	5.04e-18
Time:	00:28:00	**Log-Likelihood:**	-63.479
No. Observations:	31	**AIC:**	129.0
Df Residuals:	30	**BIC:**	130.4
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
FlightLength	0.0552	0.003	18.606	0.000	0.049	0.061

Omnibus:	2.643	**Durbin-Watson:**	1.424
Prob(Omnibus):	0.267	**Jarque-Bera (JB):**	2.221
Skew:	-0.542	**Prob(JB):**	0.329
Kurtosis:	2.263	**Cond. No.**	1.00

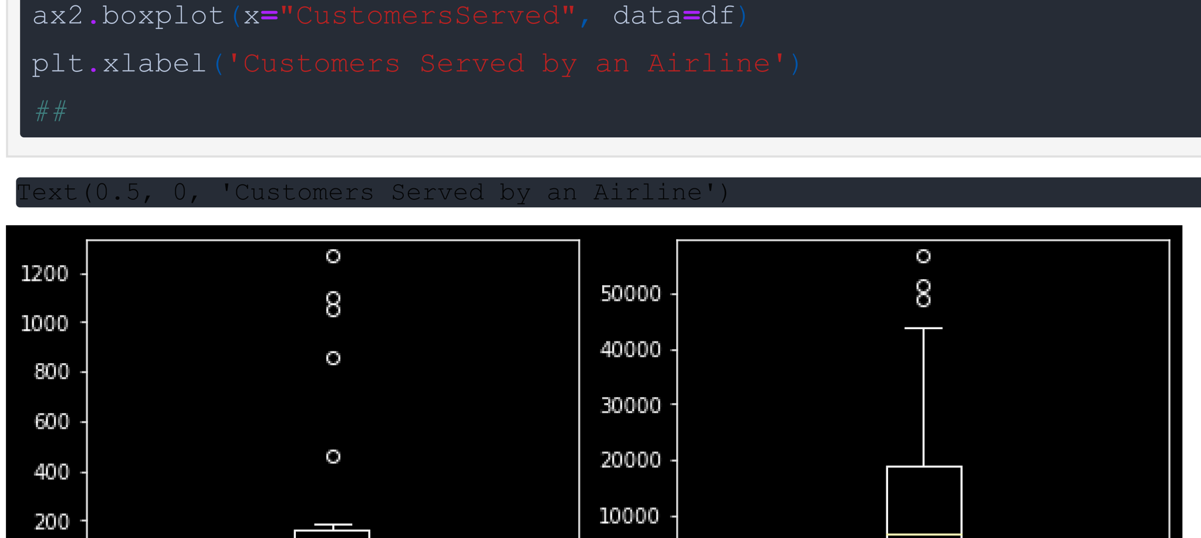
Notes:
[1] R^2 is computed without centering (uncentered) since the model does not contain a constant.
[2] Standard Errors assume that the covariance matrix of the errors is correctly specified.

1.2

The second regression model was conducted to discover the relationship between total assets by an airline (AdjustedAssets) and customers served by an airline (CustomersServed). The initial exploratory analyses were conducted to detect potential outliers. There were various outliers discovered in both columns. The identical approach was taken as the prior regression model regarding replacing the potential outliers with their respective median.

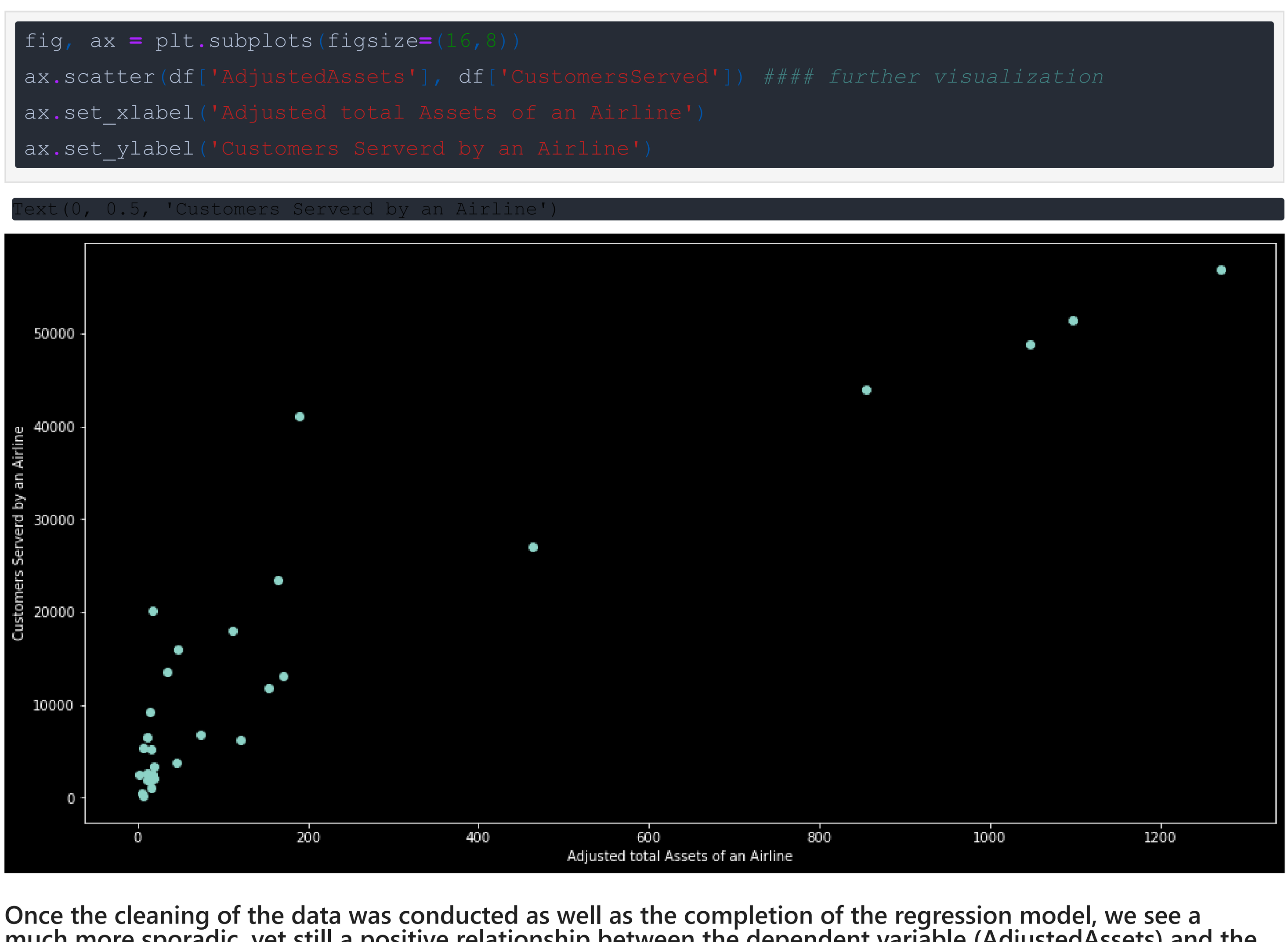
In [22]:
```python
fig = plt.figure(figsize = (15,7)) ### identifying potential outliers with boxplots
plt.style.use('dark_background')
ax1 = fig.add_subplot(231)
ax1.boxplot(x="AdjustedAssets", data=df)
plt.xlabel('Adjusted total Assets of an Airline')
##
ax2 = fig.add_subplot(232) ### various potential outliers detected
ax2.boxplot(x="CustomersServed", data=df)
plt.xlabel('Customers Served by an Airline')
##
```

Out[22]:
```
Text(0.5, 0, 'Customers Served by an Airline')
```



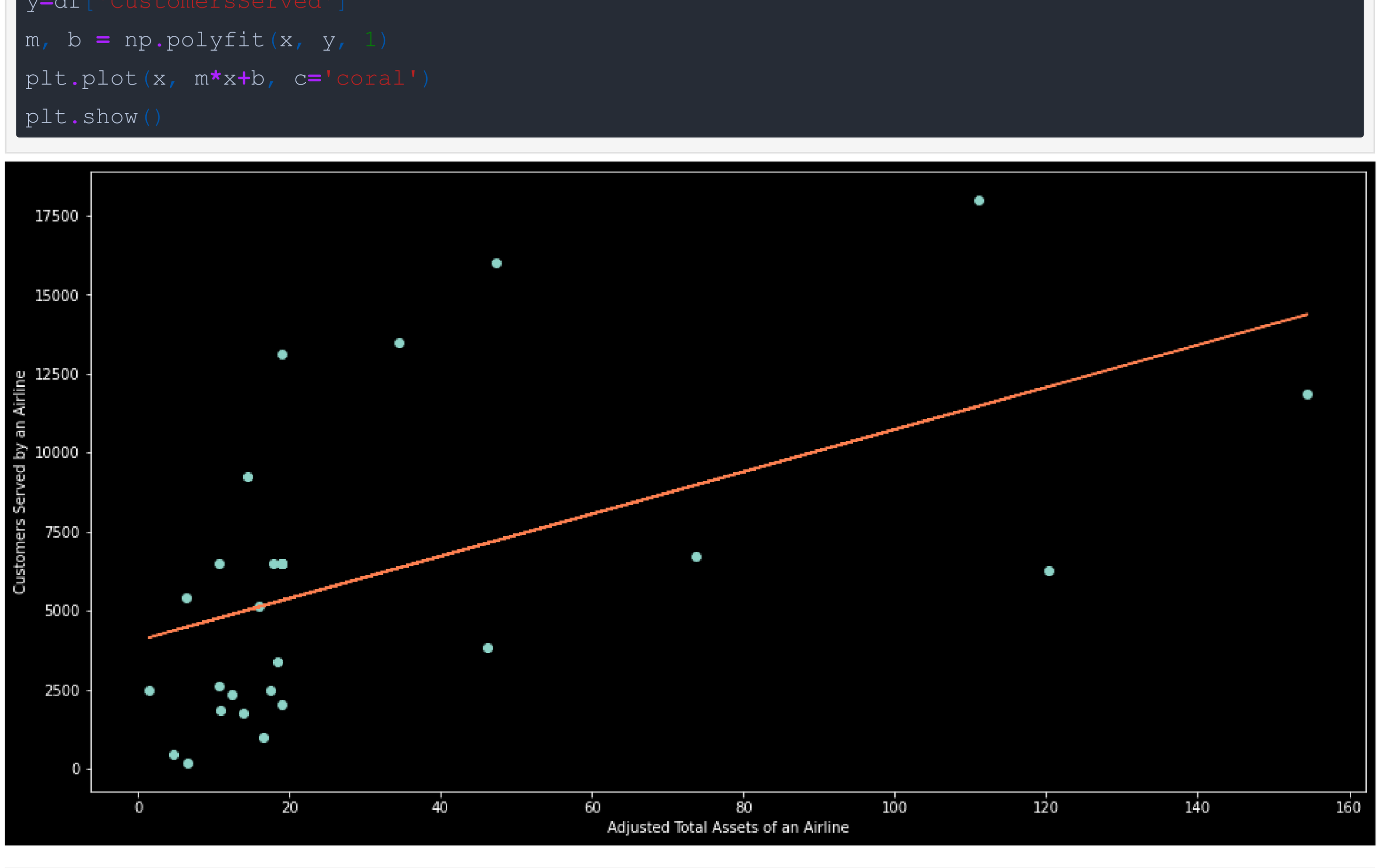
In [23]:
```python
fig, ax = plt.subplots(figsize=(16,8))
ax.scatter(df['AdjustedAssets'], df['CustomersServed']) #### further visualization
ax.set_xlabel('Adjusted total Assets of an Airline')
ax.set_ylabel('Customers Serverd by an Airline')
```

Out[23]:
```
Text(0, 0.5, 'Customers Serverd by an Airline')
```



Once the cleaning of the data was conducted as well as the completion of the regression model, we see a much more sporadic, yet still a positive relationship between the dependent variable (AdjustedAssets) and the predictor variable (CustomersServed). The R-squared value is much lower than the first regression model, being reported at 58% (.586). The coefficient for the dependent variable is also reported to be at 123.08. This indicates that as the dependent variable increased by 1, the predictor variable increases by approximately 123.08, in a positive, yet sporadic, relationship.

In [26]:
```python
df['AdjustedAssets'] =df['AdjustedAssets'].mask(df['AdjustedAssets'] > 159.8100, median)
df['CustomersServed'] =df['CustomersServed'].mask(df['CustomersServed'] > 19100, median)
### replacing potential outliers with their respective medians
fig, ax = plt.subplots(figsize=(16,8))
ax.scatter(df['AdjustedAssets'], df['CustomersServed']) #### further visualization
ax.set_xlabel('Adjusted Total Assets of an Airline')
ax.set_ylabel('Customers Served by an Airline')
x=df['AdjustedAssets']
y=df['CustomersServed']
m, b = np.polyfit(x, y, 1)
plt.plot(x, m*x+b, c='coral')
plt.show()
```



In [28]:
```python
x = df['AdjustedAssets']
y = df['CustomersServed']
mo2 = sm.OLS(y, x).fit()
predictions2 = mo2.predict(x)
mo2.summary()
```

Out[28]:

OLS Regression Results

Dep. Variable:	CustomersServed	**R-squared (uncentered):**	0.586
Model:	OLS	**Adj. R-squared (uncentered):**	0.573
Method:	Least Squares	**F-statistic:**	42.53
Date:	Tue, 08 Dec 2020	**Prob (F-statistic):**	3.28e-07
Time:	00:31:35	**Log-Likelihood:**	-306.90
No. Observations:	31	**AIC:**	615.8
Df Residuals:	30	**BIC:**	617.2
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
AdjustedAssets	123.0879	18.873	6.522	0.000	84.543	161.632

Omnibus:	2.356	**Durbin-Watson:**	1.451
Prob(Omnibus):	0.308	**Jarque-Bera (JB):**	1.151
Skew:	-0.357	**Prob(JB):**	0.562
Kurtosis:	3.618	**Cond. No.**	1.00

Notes:
[1] R^2 is computed without centering (uncentered) since the model does not contain a constant.
[2] Standard Errors assume that the covariance matrix of the errors is correctly specified.

In conclusion of the two regression models that were built, it can be concluded that both models experienced positive linear relationships. However, the first regression model had more of a consistant and accurate conclusion due to it's R-squared value being confidently high in regard to the second regression model. There is accurate and statistically significant data present in both models due to their high F-Statistic as well. The Standard Error is also considerably low in the first model (0.03), further proving the validity that the data competently represents the sample of the overall population. Even though the Standard Error for the second model is considerably higher, (18.87) there still is statistical significance due to the high F-Statistic. Therefore, it is viable that the longer the flights, the more customers are being served by each airline. However, even though there is a positive relationship between the total assets of an airline from the customers served by each airline, there is a much higher margin for error of the data's validity.