DATS 6313 – Time Series Analysis & Modeling
Instructor: Reza Jafari
Lab #1
Bradley Reardon
1/21/2022

1 – Abstract:

This lab pertains to using various visual and statistical methods for determining if a subset of data is stationary. The methods used throughout this lab are visualizing data using graphs and using both an ADF-test and KPPS-test. The results and observations of all three methods are compared to determine if they reinforce one another.

2 – Introduction:

This experiment was performed to increase understanding of stationary versus non-stationary data, and how to test for such using both visual and statistical approaches. Stationary data has a consistent mean, variance, and autocorrelation structure that do not change over time, meaning they metrics are constant throughout the subset of data being tested. As mentioned in the abstract, some methods for testing if a subset of data is stationary are visual cues, ADF-tests, and KPPS-tests.

3 – Method, Theory, and Procedures:

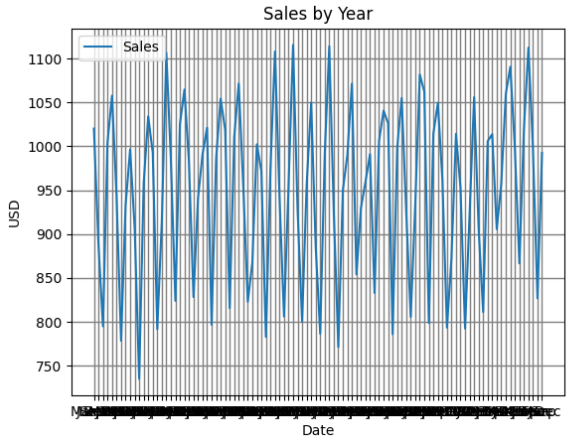
The first method we used was checking visual cues. To check for stationary data using a visual approach, we calculate the rolling mean and variance of a subset of data and plot it on a graph to see if the trend spikes at all or if it remains constant throughout. The rolling metrics are calculated by finding the difference between specific value and a value located at a designated interval from said specific value, and this is repeated for each sample in the subset of data.

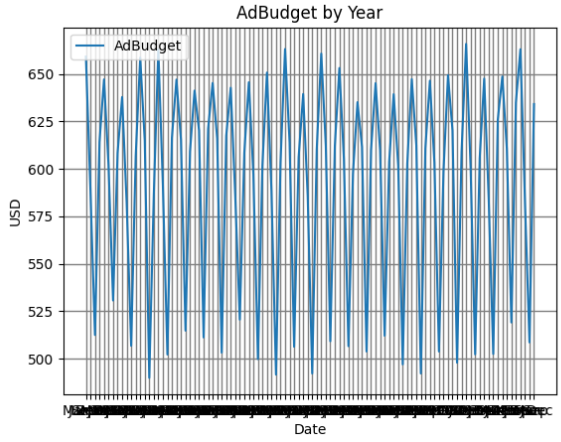
The next method used was the ADF-test, which is a test to determine if a unit root is present in a time series sample, or also known as determining if the data is stationary. The ADF-test produces results that can be used inform the degree to which a null hypothesis can be rejected or failed to be rejected. If the null hypothesis is failed to be rejected, it suggests the time-series has a unit root and is non-stationary. If the null hypothesis is rejected, it suggests that the time series is stationary. The resulting p-value will help us determine whether to reject or not reject the null hypothesis, usually at a 5% threshold while working with a 95% confidence interval. Below the p-value threshold suggests we reject, and vice versa.

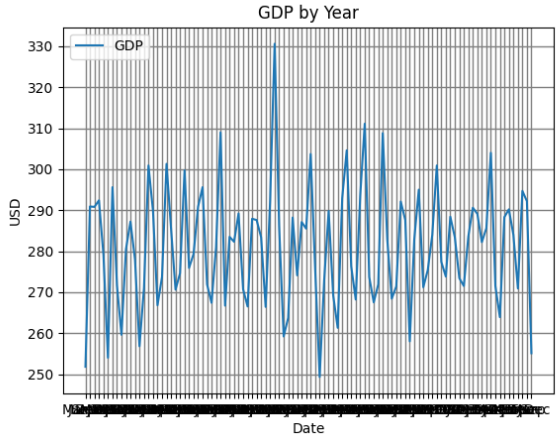
The final method used was the KPPS-test. Similarly, to the ADF-test, this is a statistical test used to determine if a time-series is stationary. One key difference between the two tests is that the null and alternative hypotheses are switched – with the KPPS test, null says the trend is stationary and the alternative says otherwise. Again, the p-value is used to determine whether to reject or fail to reject the null hypothesis and the criteria is the same as with the ADF-test.

4 – Answers to Lab Questions:

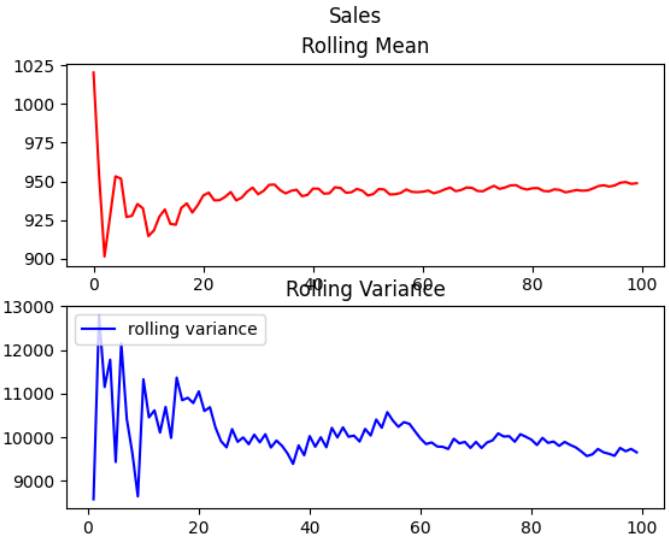
1.



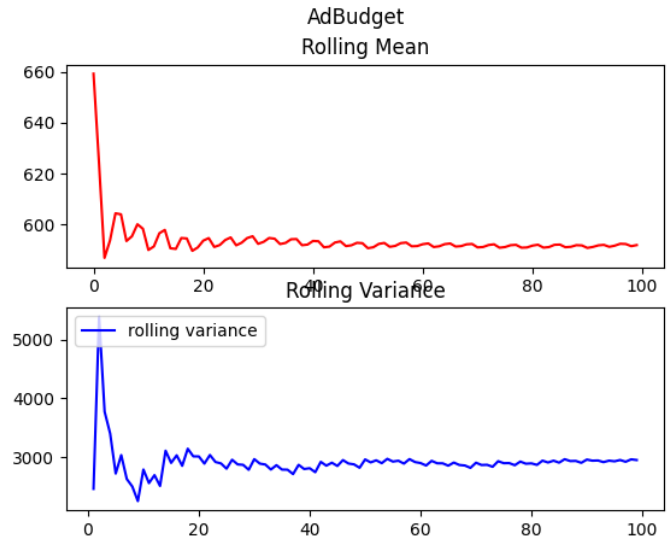


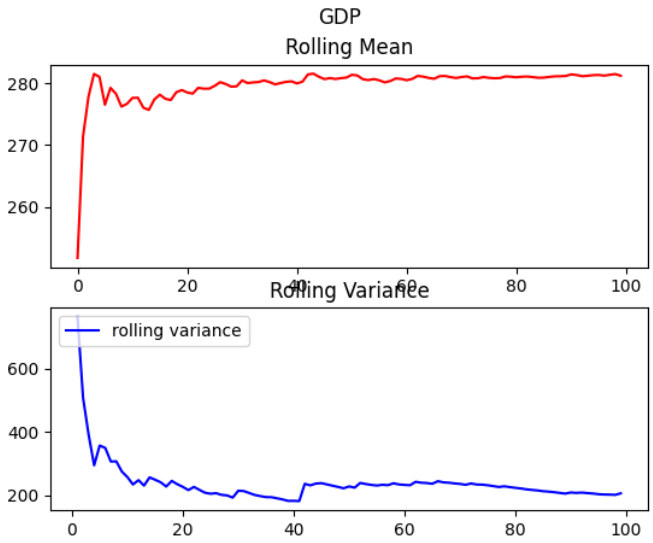


2. The Sales mean is : 948.737 and the variance is : 9653.492253535353 with standard deviation : 98.2521870165512
   The AdBudget mean is : 591.933 and the variance is : 2953.103041414141 with standard deviation : 54.34246075965038
   The GDP mean is : 281.18300000000005 and the variance is : 206.51092020202032 with standard deviation : 14.370487820600257


Sales
Rolling Mean
1025
1000
975
950
925
900
0
20
40
60
80
100
Rolling Variance
rolling variance
13000
12000
11000
10000
9000
0
20
40
60
80
100

3.


AdBudget
Rolling Mean
660
640
620
600
0
20
40
60
80
100
Rolling Variance
rolling variance
5000
4000
3000
0
20
40
60
80
100



4.

Sales:
The rolling mean begins to stabilize, but still shows some variance as samples are included. I would consider it to be stationary as samples are added. The rolling variance seems to be unstable all the way through each additional sample, rendering it non-stationary.

ADBudget:
Both rolling mean and rolling variance seem to stabilize early on in the addition of samples and seem to both become stationary as there is little variance in the rolling values from then on.

GDP:
The rolling mean and variance both become stable with little variance about halfway through the addition of samples, but are not stable preceding the halfway point and have high variance with the addition of each sample. This shows being non-stationary with the first half of the samples, but stationary with the addition of the latter half of samples.

5.

```
Sales: p-value <5%; reject null hypothesis. Assume no root (data is not
stationary)
AdBudget: p-value >5%; fail to reject null hypothesis. Assume root
(data is stationary)
GDP: p-value <5%; reject null hypothesis. Assume no root (data is not
stationary)

The ADF test reinforces my observations in question 4.

--Sales ADF--
ADF Statistic: -3.262755
```

```
p-value: 0.016628
Critical Values:
  1%: -3.505
  5%: -2.894
  10%: -2.584
--AdBudget ADF--
ADF Statistic: -2.758605
p-value: 0.064434
Critical Values:
  1%: -3.504
  5%: -2.894
  10%: -2.584
--GDP ADF--
ADF Statistic: -3.227577
p-value: 0.018443
Critical Values:
  1%: -3.504
  5%: -2.894
  10%: -2.584
```

6.

```
The test statistics for Sales, AdBudget, and GDP are all lower than the
critical value given
a confidence interval of 95% (critical value 5%), which aligns with the
p-value for each
subset of data being >0.05. This means we fail to reject the null
hypothesis for each
subset of data, making the assumption that all three subsets of data
are stationary.

The results of the kpss do not reinforce the observations of the
previous steps for
the Sale and GDP subsets, but does for the AdBudget subset.

--Sales kpss--
Results of KPSS Test:
Test Statistic            0.305544
p-value                   0.100000
LagsUsed                 19.000000
Critical Value (10%)      0.347000
Critical Value (5%)       0.463000
Critical Value (2.5%)     0.574000
Critical Value (1%)       0.739000
dtype: float64
--AdBudget kpss--
Results of KPSS Test:
Test Statistic            0.087946
p-value                   0.100000
LagsUsed                 14.000000
Critical Value (10%)      0.347000
Critical Value (5%)       0.463000
Critical Value (2.5%)     0.574000
Critical Value (1%)       0.739000
dtype: float64
--GDP kpss--
Results of KPSS Test:
```

```
Test Statistic            0.319751
p-value                   0.100000
LagsUsed                 42.000000
Critical Value (10%)      0.347000
Critical Value (5%)       0.463000
Critical Value (2.5%)     0.574000
Critical Value (1%)       0.739000
```

5 – Conclusion:

Whether a subset of data is stationary or not indicates if the mean, variance, and autocorrelation do not change over time. A few approaches for determining if your data is stationary include a visual check by plotting the rolling mean and variance, conducting an ADF-test, and conducting a KPPS-test. In this lab, we used all three approaches and cross referenced the results from each to see if our observations were consistent throughout. This lab involved data regarding quarterly sales for a small company over the period of 1981-2005. The features tested to determine if they were stationary are Sales, AdBudget, and GDP. I found some inconsistency amongst the features when performing the three tests, with the visual check seeming to show both Sales and GDP as non-stationary and AdBudget as stationary, the ADF-test following suit with the visual test observations, but the KPPS-test results showing all features as stationary. This proves that no single test is always representative of the actuality of stationary data, and that multiple tests should be used to cross-reference results with one another. One problem with the visual approach is that the observations are subjective and cannot always be interpreted correctly.

6 – Appendix

```
# packages
import matplotlib.pyplot as plt
import pandas as pd
from statsmodels.tsa.stattools import adfuller
from toolbox import ADF_Cal, kpss_test



# load data
'''OR_PATH = os.getcwd()
os.chdir('..')
PATH = os.getcwd()
DATA_DIR = PATH + os.path.sep + "Datasets" + os.path.sep
os.chdir(OR_PATH)'''

df = pd.read_csv(r'C:\Users\brear\OneDrive\Documents\GitHub\Time-Series-
Analysis-and-Moldeing\Datasets\tute1.csv')
df.rename(columns={'Unnamed: 0': 'Date'}, inplace=True)
print(df.head())

# Question 1
'''fig, (ax1, ax2, ax3) = plt.subplots(3)
fig.suptitle('Plots')
ax1.plot(df['Date'], df['Sales'])
ax2.plot(df['Date'], df['AdBudget'])
```

```
ax3.plot(df['Date'], df['GDP'])'''

plt.plot(df['Date'], df['Sales'], label='Sales')
plt.grid(color='gray', linestyle='-', linewidth=1)
plt.legend(loc='upper left')
plt.title('Sales by Year')
plt.xlabel('Date')
plt.ylabel('USD')
plt.show()

plt.plot(df['Date'], df['AdBudget'], label='AdBudget')
plt.grid(color='gray', linestyle='-', linewidth=1)
plt.legend(loc='upper left')
plt.title('AdBudget by Year')
plt.xlabel('Date')
plt.ylabel('USD')
plt.show()

plt.plot(df['Date'], df['GDP'], label='GDP')
plt.grid(color='gray', linestyle='-', linewidth=1)
plt.legend(loc='upper left')
plt.title('GDP by Year')
plt.xlabel('Date')
plt.ylabel('USD')
plt.show()

print('---------Question 2---------')
sales_mean = df['Sales'].mean()
sales_var = df['Sales'].var()
sales_std = df['Sales'].std()
print('The Sales mean is : {} and the variance is : {} with standard
deviation : {}'.format(sales_mean, sales_var, sales_std))

adb_mean = df['AdBudget'].mean()
adb_var = df['AdBudget'].var()
adb_std = df['AdBudget'].std()
print('The AdBudget mean is : {} and the variance is : {} with standard
deviation : {}'.format(adb_mean, adb_var, adb_std))

gdp_mean = df['GDP'].mean()
gdp_var = df['GDP'].var()
gdp_std = df['GDP'].std()
print('The GDP mean is : {} and the variance is : {} with standard deviation
: {}'.format(gdp_mean, gdp_var, gdp_std))

# Question 3

#  Sales
sales_rolling_mean = []
for i in range(len(df['Sales'])):
    iter_mean = df['Sales'].loc[:i].mean()
    sales_rolling_mean.append(iter_mean)
# print(sales_rolling_mean)

sales_rolling_var = []
for i in range(len(df['Sales'])):
    iter_var = df['Sales'].loc[:i].var()
```

```
    sales_rolling_var.append(iter_var)
# print(sales_rolling_var)

fig_sales, (ax1, ax2) = plt.subplots(2)
fig_sales.suptitle('Sales')
ax1.plot(sales_rolling_mean, label='rolling mean', c='r')
plt.legend(loc='upper left')
ax1.set_title('Rolling Mean')
ax2.plot(sales_rolling_var, label='rolling variance', c='b')
ax2.set_title('Rolling Variance')
plt.tight_layout
plt.legend(loc='upper left')
plt.show()

#  ADBudget
adb_rolling_mean = []
for i in range(len(df['AdBudget'])):
    iter_mean = df['AdBudget'].loc[:i].mean()
    adb_rolling_mean.append(iter_mean)
# print(abd_rolling_mean)

adb_rolling_var = []
for i in range(len(df['AdBudget'])):
    iter_var = df['AdBudget'].loc[:i].var()
    adb_rolling_var.append(iter_var)
# print(abd_rolling_var)

fig_sales, (ax1, ax2) = plt.subplots(2)
fig_sales.suptitle('AdBudget')
ax1.plot(adb_rolling_mean, label='rolling mean', c='r')
plt.legend(loc='upper left')
ax1.set_title('Rolling Mean')
ax2.plot(adb_rolling_var, label='rolling variance', c='b')
ax2.set_title('Rolling Variance')
plt.tight_layout
plt.legend(loc='upper left')
plt.show()

#  GDP
gdp_rolling_mean = []
for i in range(len(df['GDP'])):
    iter_mean = df['GDP'].loc[:i].mean()
    gdp_rolling_mean.append(iter_mean)
# print(gdp_rolling_mean)

gdp_rolling_var = []
for i in range(len(df['GDP'])):
    iter_var = df['GDP'].loc[:i].var()
    gdp_rolling_var.append(iter_var)
# print(gdp_rolling_var)

fig_sales, (ax1, ax2) = plt.subplots(2)
fig_sales.suptitle('GDP')
ax1.plot(gdp_rolling_mean, label='rolling mean', c='r')
plt.legend(loc='upper left')
ax1.set_title('Rolling Mean')
ax2.plot(gdp_rolling_var, label='rolling variance', c='b')
```

```
ax2.set_title('Rolling Variance')
plt.tight_layout
plt.legend(loc='upper left')
plt.show()

# Question 4
print('---------Question 4---------')
'''
Sales:
The rolling mean begins to stabilize, but still shows
some variance as samples are included. I would consider it
to be stationary as samples are added. The rolling variance seems
to be unstable all the way through each additional sample,
rendering it non-stationary.

ADBudget:
Both rolling mean and rolling variance seem to stabilize
early on in the addition of samples and seem to both become stationary
as there is little variance in the rolling values from then on.

GDP:
The rolling mean and variance both become stable with little variance
about halfway through the addition of samples, but are not stable
preceding the halfway point and have high variance with the addition of
each sample. This shows being non-stationary with the first half of the
samples, but stationary with the addition of the latter half of samples.

'''

# Question 5
print('---------Question 5---------')
print('--Sales ADF--')
ADF_Cal(df['Sales'])

print('--AdBudget ADF--')
ADF_Cal(df['AdBudget'])

print('--GDP ADF--')
ADF_Cal(df['GDP'])

'''
Sales: p-value <5%; reject null hypothesis. Assume no root (data is not
stationary)
AdBudget: p-value >5%; fail to reject null hypothesis. Assume root (data is
stationary)
GDP: p-value <5%; reject null hypothesis. Assume no root (data is not
stationary)

The ADF test reinforces my observations in question 4.
'''

# Question 6
print('---------Question 6---------')
print('--Sales kpss--')
kpss_test(df['Sales'])

print('--AdBudget kpss--')
```

```
kpss_test(df['AdBudget'])

print('--GDP kpss--')
kpss_test(df['GDP'])

'''
The test statistics for Sales, AdBudget, and GDP are all lower than the
critical value given
a confidence interval of 95% (critical value 5%), which aligns with the p-
value for each
subset of data being >0.05. This means we fail to reject the null hypothesis
for each
subset of data, making the assumption that all three subsets of data are
stationary.

The results of the kpss do not reinforce the observations of the previous
steps for
the Sale and GDP subsets, but does for the AdBudget subset.
'''
```