The George Washington University

Laboratory One:

Describing and Visualizing Stationary Data

Fernando Zambrano

DATS 6450: Multivariate Modeling

Dr. Reza Jafari

29 January 2020

Abstract:

Discover the initial steps of time-series analysis using a small company's monthly financial data. Strong emphasis on visualizing time related patterns in the data. Demonstrating the differences between stationery and nom-stationery data. Explore methods and tests to determine when data is stationary or non-stationery.

Introduction:

When facing a data driven research problem or project, understanding and understanding how to describe that data is fundamental. Knowing what patterns and relationships are within the data will allow give insights into what kind of analysis can be performed, or what further data augmentation must be done to get the data in the right format for desired analyses. Undermining the importance of this initial exploration process can lead to incorrect results and insights which is accompanied by unnecessary time and effort to correct such results. In the realm of time-series analysis, which is centered on forecasting and predictive modeling, must also undergo the same initial processes. Time-series analysis is a popular and important form a data analysis since a vast majority of phenomena are naturally related and dependent on time. However, a common mistake is to perform regression, plot the line-of-best-fit, and extrapolate from these insights to predict the future on any dataset that has time as a variable or feature. This will lead to a shallow interpretation of the data with grossly inaccurate results. Instead, a more subtle pattern must discover within the data, whether it is stationary or non-stationary. Stationary data bears the possibility of a fruitful forecasting model. This can be determined by visual interpretation or by applying a more statistically objective method, the Augmented Dickey-Fuller (ADF) test. A non-stationary dataset results in greater difficulty in producing a reliable forecasting model. In the case of data being non-stationary, there are methods to transform it into stationary, but are beyond the scope of this study. Instead, the purpose of this study is to demonstrate crucial early steps required to determine if a successful time-series analysis can be performed. This will be done by analyzing quarterly financial data (sales, advertising budget, and GDP) of small company over the period of 1981-2005 as a case study.

Theory and Methods:

Data that is observed in a sequential order over time is considered time-series data. A time-series dataset can be designated as either stationary or non-stationary. The difference between the two groups is how their statistics (mean, variance, covariances) behave over time. Datasets are classified as stationary when its statistics act the same way over time so that a pattern can be recognized. It does not mean that the series does not change over time, but rather the way it changes does not vary over time. This pattern gives insight into how future observations should fit given this past data. This allows designing a more reliable predicative model given the consistency in the statistics. Non-stationary datasets are determined if its statistics behave in an inconsistent and unpredictable manner. In other words, unlike stationary data which does not vary in the way it changes through time, non-stationary does. This naturally makes prediction more difficult as there is no reliable pattern to base a forecasting model on. Hence, if the data can be classified early on as stationary, then the process of building a forecasting model is more dependable because the features within that dataset will work well with forecasting analysis.

There are two methods that will be explored in order to distinguish whether a dataset is stationary or not. The first option is to visualize the data by plotting its features over time and looking for obvious patterns. This method is highly subjective on the readers interpretation of the data which may result in an inaccurate conclusion of stationarity. However, it offers the analyst into what kind of patterns the data may follow. There are three patterns that can help determine stationarity of the dataset: trend, seasonality, and cyclicity. Trend is the "direction" which the data is moving towards. Is it mainly increasing or decreasing over time? If there is an obvious trend, then it is likely the data is non-stationary as the mean is increasing at a varying rate over time. Seasonality is the effect of the specific points in time that change the behavior of the data. These changes happen in fixed and known frequencies which allows for predictable changes. Cyclicity refers to the data exhibiting cycle of rises and falls that occur in irregular frequencies. Cyclic behavior can be clear in hindsight but becomes more unpredictable in the long-term future. The differences between seasonal and cyclic is that seasonal changes occur at constant length, whereas changes in cyclic vary. Additionally, for reemphasis, seasonality is dependent directly with time, while cyclicity is a result of previous values. Furthermore, seasonality and cyclicity are not clear signs of the data being non-stationary, as long as the change in the variance remains constant over time the data will remain stationary.

The second option is to perform a statistical hypothesis test which determines the likelihood of the dataset being non-stationary. The Augmented Dickey-Fuller (ADF) test, or "unit root test" makes a strong assumption that dataset in hand is non-stationary. This test sets up two hypotheses:

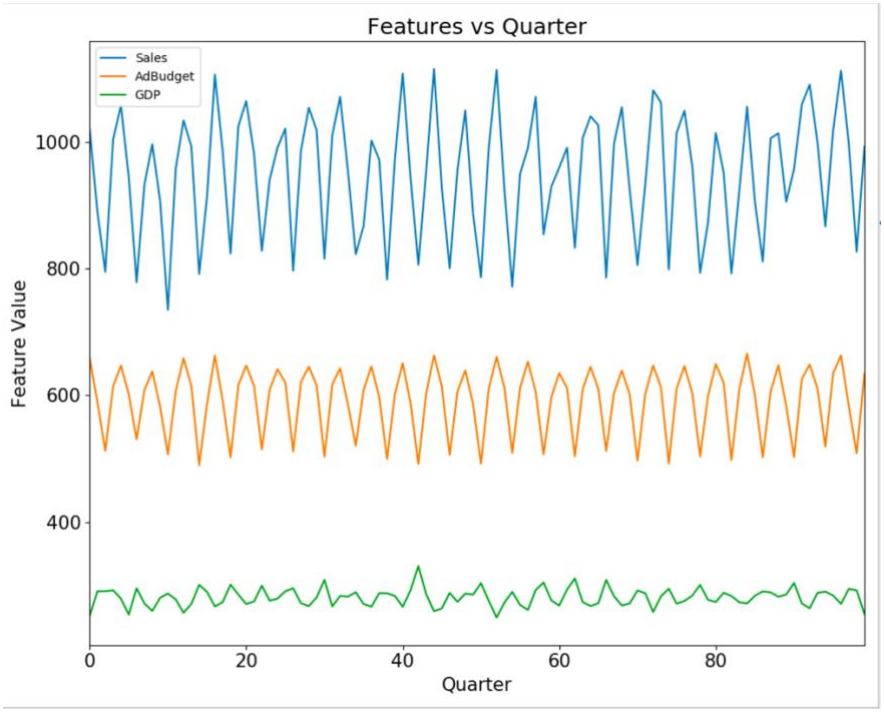
Null Hypothesis (H0): if failed to be rejected, then the time-series has a unit root and the data is non-stationary.

The Alternative Hypothesis (H1): if the null hypothesis is rejected, then there is no unit root and the data is stationary.

In order to reject the null hypothesis, there must be strong evidence to suggest that the likelihood of the data being non-stationary is significantly unlikely. This is done by setting high confidence levels which are determined by the predetermined critical values. The standard practice is to have between a 95% - 99% confidence level. A 0.05 and 0.01 critical values correspond with a 95% and 99% confidence level, respectively. The ADF test will result with a p-value which is compared to the set critical value. If the p-value is less than the critical value, then there is significantly strong evidence to suggest that the null hypothesis (non-stationary) be rejected in favor of the alternative (stationary). Conversely, if the p-value is greater than the critical value, then there is not enough significantly strong evidence to support rejecting the null hypothesis.

Implementation and Results:

The first step in the analysis is to load the small company's financial dataset and plot the individual features over time. The graph below demonstrates how sales, AdBudget (advertising budget), and company GDP (revenue) change over time. From a visual analysis, the dataset appears as stationary, as the mean and variance appear constant over time. Yet, further evidence is needed to support this claim.

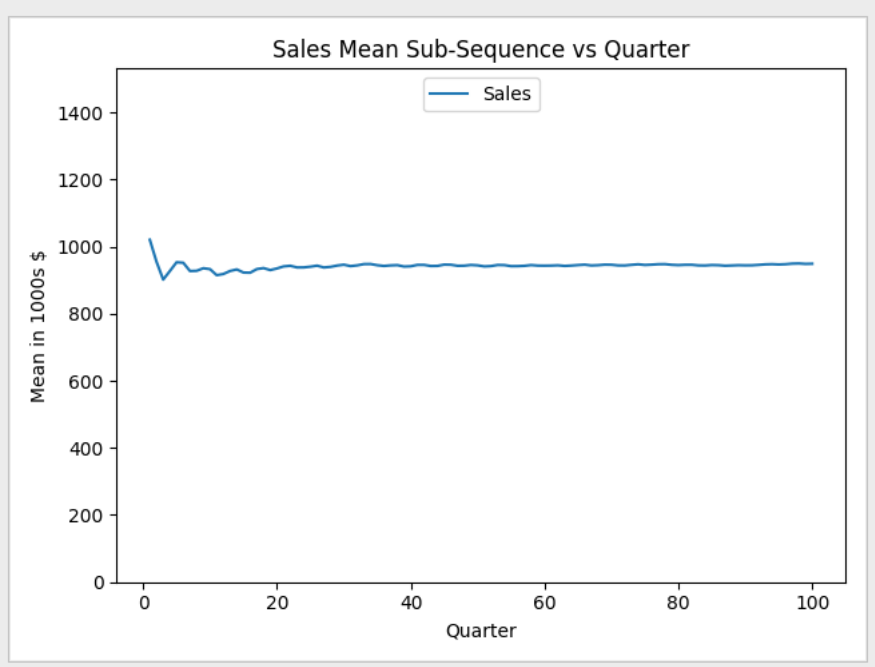


The next step is to get more statistical insight of the overall data by calculating the average, variance, and standard deviation for each feature.

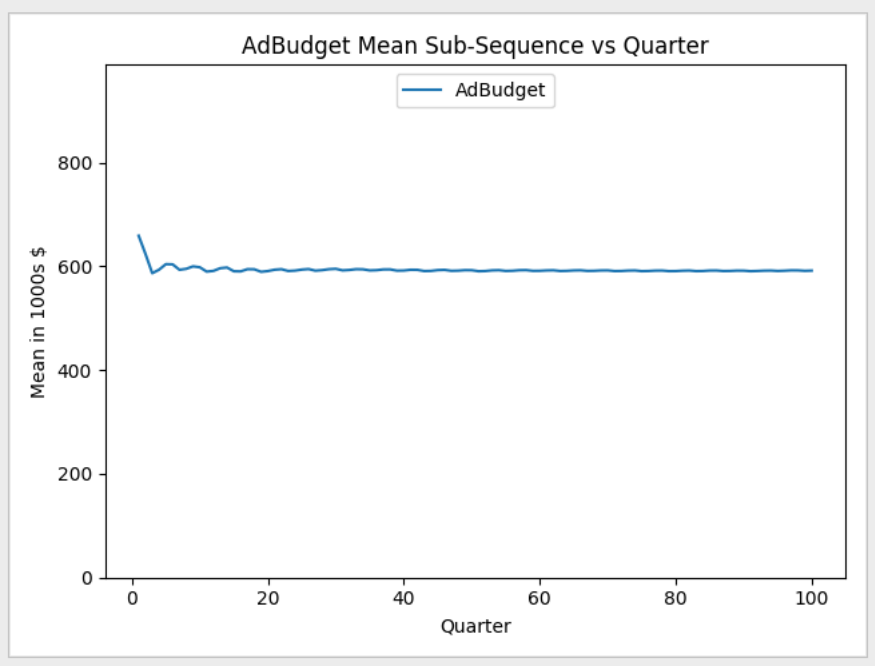
```
The Sales mean is : 948.737 and the variance is : 9653.492253535354 with s
tandard deviation : 98.252187

The AdBudget mean is : 591.933 and the variance is : 953.1030414141414 wit
h a standard deviation : 54.342461

The GDP mean is : 281.183 and the variance is : 206.5109202020203 with sta
ndard deviation : 14.370488
```

These overall statistical may be useful for describing the data and for other analysis, but they do not give insight to how the mean and variance change over time which is crucial to determine if the data is stationary or non-stationary.

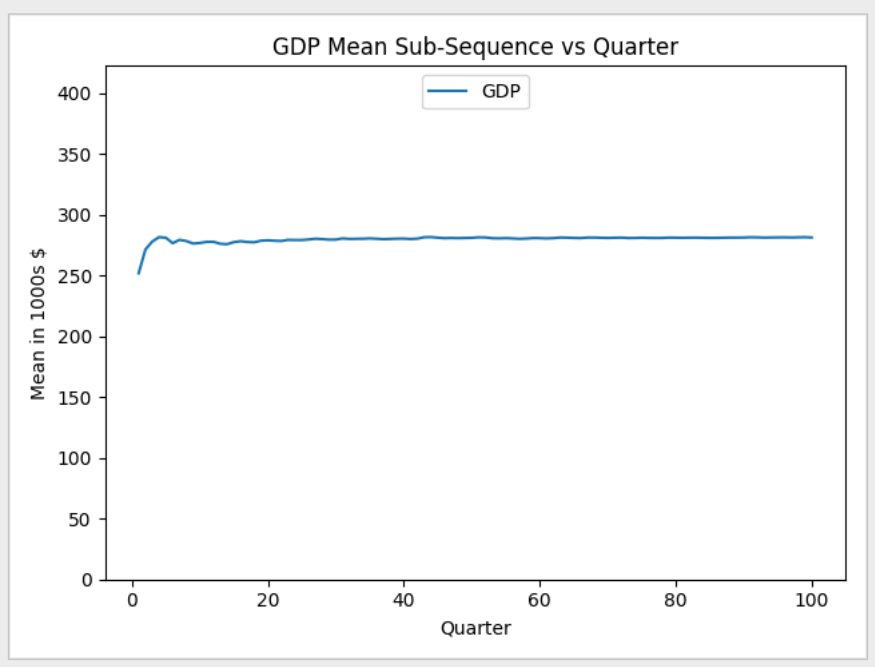
In order to support the first visual analysis that each data feature is stationary, there needs to be evidence that the mean and variance are remaining constant over time, rather than showing a clear pattern varying change. This requires sub-sequencing the data by taking the mean and variance at each quarter time-step. The following first three graphs will show the sub-sequencing of the mean for each feature, and the subsequent three graphs with show the sub-sequencing of variance. Using a combination of trend, seasonality, and cyclicity as visual criteria for classifying the feature data as stationary or non-stationary.



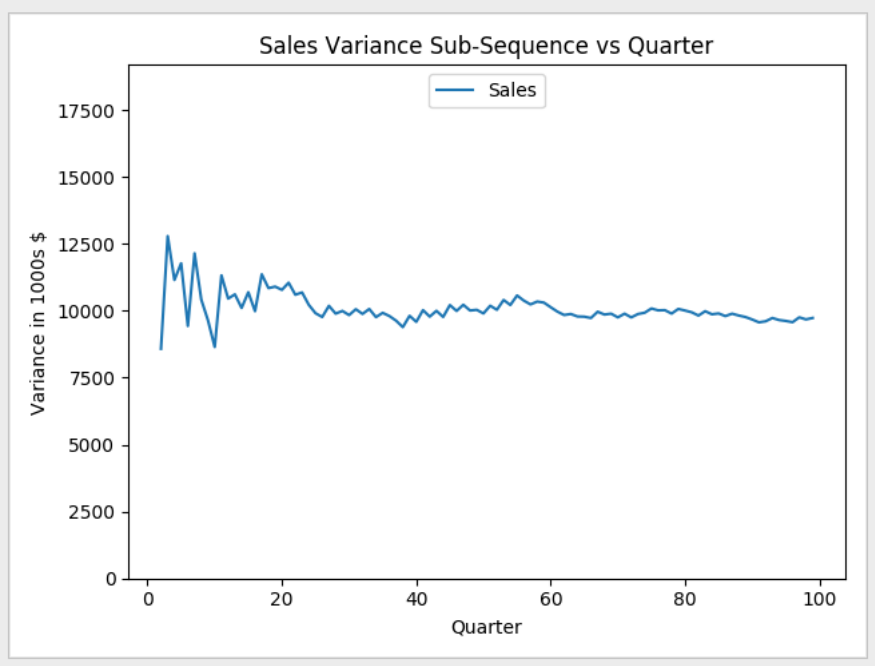
The graph demonstrates there is no clear trend in the mean which suggests that the sales data is stationary.



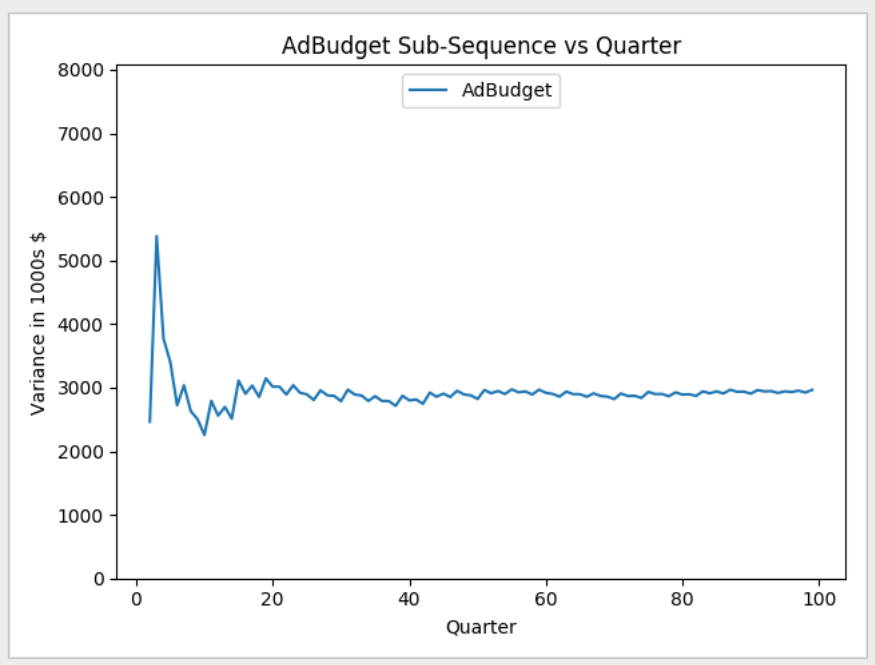
The graph demonstrates that there is no clear trend in the mean which suggests that the AdBudget data is stationary.



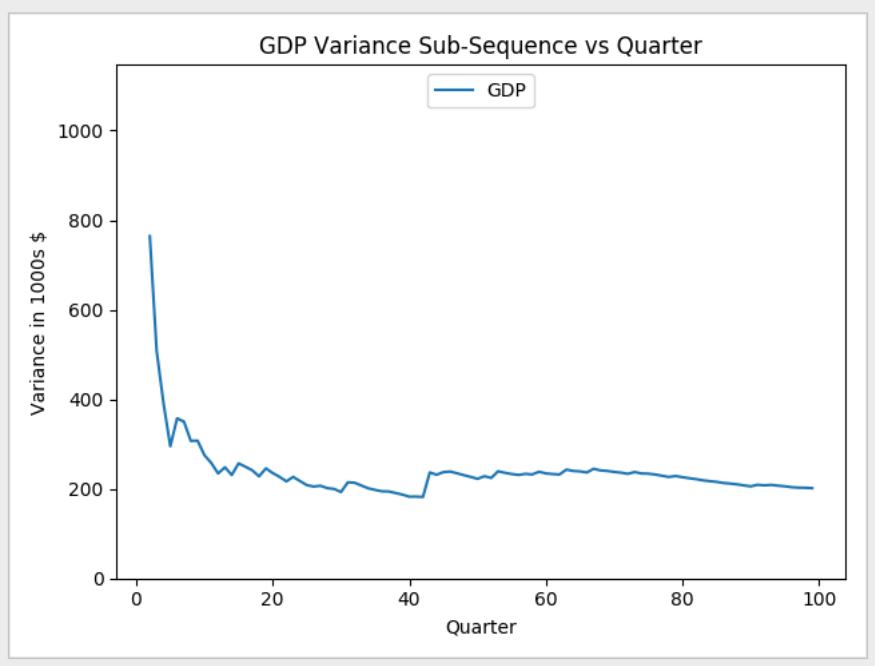
The graph demonstrates that there is no clear trend in the mean which suggests that the GDP data is stationary.



The graph demonstrates that there is no clear trend in the variance which suggests that sales data is stationary.



The graph demonstrates that there is no clear trend in the variance which suggests that AdBudget data is stationary



The graph demonstrates that there might be weak downward trend in the variance which suggests that GDP data could be non-stationary. However, my subjectivity would lean more towards stationary.

Below are the results of the ADF test on each feature. If the p-value is less than critical values (0.05) then the data is considered stationary.

Sales
ADF Statistic: -3.262755
p-value: 0.016628
Critical Values:
1%: -3.505190
5%: -2.894232
10%: -2.584210

AdBudget
ADF Statistic: -2.758605
p-value: 0.064434
Critical Values:
1%: -3.503515
5%: -2.893508
10%: -2.583824

GDP
ADF Statistic: -3.227577
p-value: 0.018443
Critical Values:
1%: -3.503515
5%: -2.893508
10%: -2.583824

The results of the ADF test demonstrate that both sales and GDP are stationary as their p-values (0.016) and (0.01), respectively are less than 0.05. However, since the p-value of AdBudget (0.0644) is slightly greater than 0.05, there is a failure to reject the null hypothesis that the data is non-stationary.

Conclusion:

Overall, this study clearly demonstrates the advantages and differences between visually or statistically analyzing the change of the mean and variance over time for each feature. Visual analysis is not always a reliable and objective manner to determine stationarity as seen with having some doubt to whether the variance of GDP was stationary. The ADF test demonstrated that my doubts about GDP data being stationary were incorrect, as it revealed strong evidence that it was stationary. Additionally, my certainty of AdBudget data being stationary were also challenged, as the ADF demonstrated that is more likely to be non-stationary. Visual interpretation of the data would be more difficult to have a definitive opinion of the data's stationarity if there was more ambiguity of the absence of a clear trend in either the mean or variance. Hence, this illustrates the importance of applying an ADF test on each feature to have a better understanding of its stationarity.

Appendix:

```
import matplotlib.pyplot as plt
import pandas as pd
import numpy as np
from statsmodels.tsa.stattools import adfuller
import os
import statistics

os.listdir()

# 1 Load the Data tute1.csv
df = pd.read_csv("tute1.csv")

# 4 Plot sales AdBudget, and GDP

ax = df[['Sales','AdBudget',"GDP"]].plot(kind='line',figsize=(10,8), fontsize=15)
plt.legend(loc='upper left', fontsize=10)
ax.set_xlabel('Time Step', fontsize=15)
ax.set_ylabel('Feature Value', fontsize=15)
ax.set_title('Features vs Time Step',fontsize=18)
plt.show()

# 5 Find the time sereis statistics for mean, variance, standard deviation
df.describe()

print("Sales Variance :",statistics.variance(df.Sales))
print("AdBudget Variance :",statistics.variance(df.AdBudget))
print("GDP Variance :",statistics.variance(df.GDP))


# 6 Display average variance and standard deviation as "feat" is: etc.

print("The Sales mean is :",948.737000,"and the variance is : ",9653.492253535354,"with standard deviation
:",98.252187)
print("The AdBudget mean is :",591.933000,"and the variance is :",953.1030414141414,"with a standard deviation
```

```
:",54.342461)
print("The GDP mean is :",281.183000,"and the variance is :",206.5109202020203,"with standard deviation
:",14.370488)


# 7 show that the features are stationary


# Function to get cumulative sum at each element stage
def summation(feature):
    dfl = feature
    summation = []
    total = 0
    for i in dfl:
        total += i
        summation.append(total)
    return summation
# Function to get the average at each element stage
def means(feature):
    agg = feature
    means = []
    for i in list(range(1,101)):
        x = agg[i-1]/i
        means.append(x)
    return means
# Functionto plot the change in statstic with each time step
def plots_mean(stat,title,feature):
    x = range(1,101)
    line = plt.plot(x,stat,label = str(feature))
    plt.xlabel("Quarter")
    plt.ylabel("Mean in 1000s $")
    plt.legend(loc="upper center")
    plt.title(str(title))
    plt.ylim(0,1.5*max(stat))
    return plt.show()
# Functionto plot the change in statstic with each time step
def plots_var(stat,title,feature):
    x = range(2,100)
```

```
    line = plt.plot(x, stat, label=str(feature))
    plt.xlabel("Quarter")
    plt.ylabel("Variance in 1000s $")
    plt.legend(loc = "upper center")
    plt.title(str(title))
    plt.ylim(0, 1.5 * max(stat))
    return plt.show()

# Plot all means and variance.
# Write down your observation about if this time series deate is stationary or not? Why?

# Calculate and visualize for mean Sales
sums = summation(df.Sales)
avg_s = means(sums)
plots_mean(avg_s,"Sales Mean Sub-Sequence vs Quarter","Sales")

# Calculate and visualize for mean AdBudget
sumad = summation(df.AdBudget)
avg_ad = means(sumad)
plots_mean(avg_ad,"AdBudget Mean Sub-Sequence vs Quarter","AdBudget")

# Calculate and visualize for mean GDP
sumgdp = summation(df.GDP)
avg_gdp = means(sumgdp)
plots_mean(avg_gdp,"GDP Mean Sub-Sequence vs Quarter","GDP")

# Plot Variances
# Take in data frame
# pull each element and add to empty list
# run variance on list
# append the resulting variance into a new list
sales_var = []
adbug_var = []
gdp_var = []
for i in range(2,100):
    sales_var.append(df.Sales.head(i).var())
    adbug_var.append(df.AdBudget.head(i).var())
```

```
    gdp_var.append(df.GDP.head(i).var())

plots_var(sales_var,"Sales Variance Sub-Sequence vs Quarter","Sales")
plots_var(adbug_var,"AdBudget Sub-Sequence vs Quarter","AdBudget")
plots_var(gdp_var,"GDP Variance Sub-Sequence vs Quarter","GDP")

# 9 Perform a an ADF test to check if the Sales, Adbudget and GDP are stationary or not?
# Does your answer for this question reinforce your observations in the previous steps?

def ADF_Cal(x):
    result = adfuller(x)
    print("ADF Statistic: %f" %result[0])
    print("p-value: %f" %result[1])
    print("Critical Values:")
    for key, value in result[4].items():
        print('\t%s: %3f' % (key,value))

ADF_Cal(df.Sales)
ADF_Cal(df.AdBudget)
ADF_Cal(df.GDP)
```

Reference: