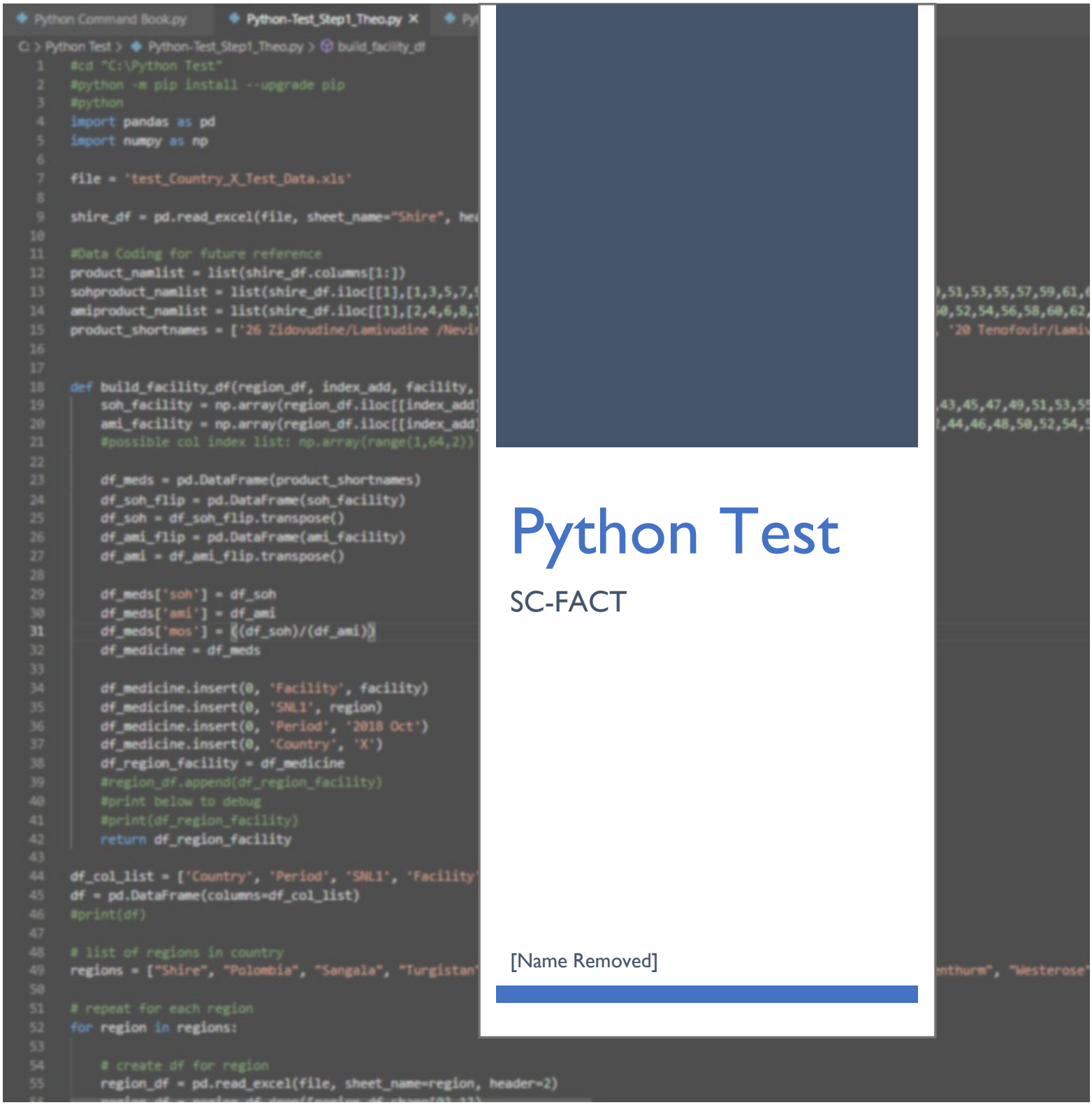

Python Test

SC-FACT

[Name Removed]

Chemonics
Development works here.

USAID
FROM THE AMERICAN PEOPLE

PEPFAR

Step 1 Global Template Excel

File: Python-Test_Step_1_[name removed].py

Commands for terminal: cd "C:\Python Test" > python Python-Test_Step_1_[name removed].py

Please watch recording titled "Python Test – Part 1 Recording - Candidate 2" to watch the code walkthrough for Step 1.

Step 2 Descriptive Statistics

The python print-out code for these descriptive statistics can be found in Python-Test_Steps_2&3_[name removed].py.

The code for each section below is included in text boxes within this document.

a. Data

The Global Template dataframe contains stock data for 32 medical products in facilities across 8 SNL1 regions. Specifically, the data measured include Stock on Hand (SOH) and Average Monthly Issuance (AMI), and we calculated the Month of Stock (MOS) remaining in step 1. All observations were collected in Country X, during the period of October 2018. The SNL1 regions include Shire, Polombia, Sangala, Turgistan, Urkesh, Nuku'la Atoll, Molvania, Hogwarts, Faulsenthurm, and Westerose. Each region has a different number of facilities, but each region tracks the same 32 medical products.

For the descriptive and inferential statistics, it makes sense to focus on MOS statistics for each region, because insights from this data can support better stock management across the country. By understanding the average levels of monthly stock remaining in each region, country X can not only identify which regions are at imminent risks of stock-outs, but also understand how to better support and improve monthly stock levels for struggling facilities and regions.

b. Measures of center by region

Python Script Lines #16-74	file = 'Global Template.xlsx' df = pd.read_excel(file) df = df.replace([np.inf], 0) df.sort_values(by='MOS', ascending=False) df_snl1 = df.loc[(df['SNL1'] == 'region name')] snl1_mos_array = np.array(df_snl1['MOS']) snl1_mos = "Average MOS: " + str(np.mean(snl1_mos_array)) + "; Median MOS: " + str(np.median(snl1_mos_array)) + "; lowest MOS: " + str(np.min(snl1_mos_array)) + "; highest MOS: " + str(np.max(snl1_mos_array)) + "; standard deviation of MOS: " + str(np.std(snl1_mos_array)) print(snl1_mos)

i. Mean MOS by region

- Shire: 4.15
- Polombia: 2.99
- Sangala: 3.56
- Turgistan: 2.97
- Urkesh: 3.95
- Nuku'la Atoll: 3.16
- Molvanîa: 8.69
- Hogwarts: 5.52
- Flausenthurm: 4.52
- Westerose: 3.51

- Country-wide: 3.98

These are average monthly stock remaining for all products by respective region. These statistics suggest that on average, each region has three months or more of monthly stocks remaining for all products. It also suggests that some regions have many more months of stock remaining for all products on average than other regions, like the 5-month stock-level difference between Molvanîa and Polombia. However, the mean can be a deceiving statistic on its own; it is important to also get the median MOS value for each region, to better understand the distribution of the observations.

ii. Median MOS by region

- Shire: 1.026
- Polombia: 1.005
- Sangala: 1.083
- Turgistan: 1.034
- Urkesh: 1.035
- Nuku'la Atoll: 0.983
- Molvanîa: 0.933
- Hogwarts: 1.022
- Flausenthurm: 1.003
- Westerose: 1.040

- Country-wide: 1.026

These statistics tell us that half of the MOS observations for each respective region fall below the median value, and the other half of the MOS observations fall above the median value. For regions with a median MOS of 0.99 and below, it means that half of the region's medical products have less than a month of stock remaining. This is a relevant statistic to have, because it gives a better idea of which regions are struggling with the stock levels of most of their products. The

median value also tells us that the mean values are affected by a heavy skew of the distribution.

It is obvious that the region most affected by outliers is Molvanîa. The average months of stock remaining for all products in Molvanîa is above 8, however, more than half of the products have less than a month of supply left. This is good information on the data's distribution, because the mean was initially deceiving; the distribution tells us that a few products in Molvanîa have an exorbitant number of months of stock remaining, but more importantly, Molvanîa is actually struggling with the stock levels for most of its products.

c. Measures of Spread

i. Standard deviation

**see code in section b.*

- Shire: 25.84
- Polombia: 11.17
- Sangala: 15.94
- Turgistan: 17.01
- Urkesh: 17.59
- Nuku'la Atoll: 10.03
- Molvanîa: 94.03
- Hogwarts: 34.46
- Flausenthurm: 45.08
- Westerose: 16.43

- Country-wide: 30.52

These standard deviation statistics show the extent to which the distribution of data is spread, and will also serve as our basis for identifying outliers. We can see that some regions like Molvanîa and Flausenthurm have a drastically wide range of monthly stock levels for different products and facilities; whereas regions like Polombia and Nuku'la Atoll have relatively more uniform monthly stock levels across products and facilities.

ii. Skewness and Kurtosis

Python Script Lines #77-79	print("MOS skewness: " + str(sp.stats.skew(mos_array))) print("MOS kurtosis: " + str(sp.stats.kurtosis(mos_array))) print(sp.stats.kurtosis(norm_dist))

Skewness tells us the degree to which outliers are affecting the distribution of data, and kurtosis tells us the volume of observations that are in both end tails of the distribution. However, to understand their degree, we need to compare our

MOS skewness and kurtosis values to the respective values of a normal distribution.

From the code above we find the skewness and kurtosis values of the country-wide MOS distribution to be:

- MOS Skewness: 35.12
- MOS kurtosis: 1557.001

Python Script Lines #80-83	norm_dist = np.random.normal(3.97956215880793, 30.51691636001731, 6496) print("Normal Distribution skewness: " + str(sp.stats.skew(norm_dist))) print("Normal Distribution kurtosis: " + str(sp.stats.kurtosis(norm_dist)))

Now we must set up a random sample normal distribution simulation using the same mean, standard deviation, and sample number as our country-wide MOS distribution. While running this simulation will return slightly different skewness and kurtosis values each time, they will stay in the same range even when the random sample size is increased.

A perfectly normal distribution with the same mean, standard deviation, and sample size as our dataset, should have skewness and kurtosis values around 0. Now, we can clearly see that our country-wide MOS distribution is heavily skewed, but more dramatically, the high kurtosis value tells us that tail-ends of the MOS distribution contain many, many outliers.

iii. Outliers

Given the mean is far higher than the median in all regions, we know that the MOS data is heavily skewed right, meaning that we have a lot of outliers on the right-side of the distribution. To address the effect of disproportionately high MOS values in the distribution, we can identify outliers. We can do this for country-wide MOS by determining the 3*StdDev value. Any MOS values exceeding that threshold will be deemed as outliers. The reasoning for this method is based on the 68-95-99.7 rule for optimal distributions, where 99.7% of data should fall within three standard deviation ranges of the mean. Therefore, any value exceeding that range is an outlier as it falls outside the normal distribution.

Python Script Lines #85-89	mos_stddev = "MOS standard deviation" + str(np.std(mos_array)) mos_3stddev = "MOS outlier floor: " + str(3*(np.std(mos_array))) print(mos_stddev) print(mos_3stddev)

For the whole country of X, any MOS value above the 3*StdDev value can be treated as an outlier.

- Outlier floor for MOS values in Country X: 91.55

iv. Other Notable Observations

An early potential issue with presenting the descriptive statistics was that the MOS value was calculated in Step 1, dividing SOH by AMI. This meant that for any observation where AMI was 0, those values ran into DIV/0 errors and therefore produced inf/-inf values. To solve for DIV/0, the following code replaced inf/-inf values:

- from numpy import inf
- df = df.replace([np.inf], 0)
- df.sort_values(by='MOS', ascending=False)

Furthermore, it must be noted that there are possible duplicate facilities in the regions of Turgistan and Nuku'la Atoll. In Turgistan, there are two facilities with the code-name "soap", and both have identical SOH and AMI values. Likewise, in Nuku'la Atoll there are two facilities code-named "tug" with identical SOH and AMI values.

d. Removing Outliers

Python Script Lines #91-126	mos_array = mos_array[(mos_array < 91.55)] print(mos_array)

This code removes the outliers that fall above our 3&StdDev floor of 91.55. This will help ease the weight of high MOS values on the distribution, and therefore we will get more accurate insights from Step 3.

Through this method, we were able to remove 34 outliers. To exemplify how removing outliers has affected the MOS distribution, we can see that the updated regional MOS means and standard deviations are more uniform across regions:

- Shire: 2.67 (6.77 std dev)
- Polombia: 2.53 (6.64 std dev)
- Sangala: 2.66 (6.95 std dev)
- Turgistan: 1.89 (2.76 std dev)
- Urkesh: 2.28 (5.85 std dev)
- Nuku'la Atoll: 2.47 (5.99 std dev)
- Molvanîa: 3.17 (9.26 std dev)
- Hogwarts: 2.71 (5.98 std dev)
- Flausenthurm: 2.56 (5.62 std dev)
- Westerose: 2.71 (6.63 std dev)

- Country-wide: 2.58 (6.39 std dev)

Step 3 Inferential Statistics

a. Hypothesis Testing

An interesting hypothesis would involve comparing the average stock level of a region compared to the stock levels of the rest of the country. This would provide insight on how well a region is faring to keep its stock levels up, in comparison to the rest of the country. If a region is struggling to keep its stock levels up to the levels of other regions, then country X knows where to target additional support and resources. On the other hand, if a region is doing comparatively well with many more months of stock levels than other regions in the country, then that successful region can serve as a model for other regions to follow.

Looking at the mean MOS of all products by region, it would appear that Molvanîa has superior stock levels compared to other regions in country X, whereas Turgistan has much shorter months of supply on average. But inferential statistics would allow us to explore deeper to see if this is truly the case within each regional data set, or if the mean values are deceiving, which is a strong possibility given the heavy right skew that is still present despite removing outliers.

i. T-Test: Molvanîa regional MOS versus country-wide MOS

To set this up, I would conduct a t-test. A t-test tells us how different two different groups are, by comparing not only the mean but especially comparing the overall distributions. Through this test, we can analyze the mean MOS of a region against the mean MOS of the rest of country, and determine whether that region's MOS is statistically higher/lower, or the same, as the mean MOS of other regions. Even though a mean MOS of one region might appear to be nominally higher than the other regions' MOS, we can't definitively conclude this is statistically true until we prove it with a t-test.

Based on the nominal mean of Molvanîa, we can set up the following hypothesis:

Null Hypothesis: The average MOS in Molvanîa is not significantly different than the average MOS in country X.
Alt Hypothesis: The average MOS in Molvania is significantly better than the average MOS in country X.

It appears that when aggregating all products, Molvanîa has on average about 18 more days of stock than the rest of the country. To determine significance in this difference of average MOS between Molvanîa and all other regions, we set a level of significance at 5%. The result of our test will provide a statistic and corresponding p-value. If the p-value falls above our set level of significance (in our case 5%, or 0.05), then we conclude that the null hypothesis is true and that there is no statistically significant difference between the average MOS in Molvanîa and the average MOS in country X. If the p-value falls below our set level of significance (p-value<0.05), then we reject the null hypothesis, and accept that Molvanîa has significantly higher MOS than the country average.

#132	print(sp.stats.ttest_ind(molvania_mos_array, mos_array))

Our t-test returns statistic=1.499 and p-value=0.133.
Since (p-value > 0.05), we fail to reject the null hypothesis, and accept that on average, Molvanîa does not have a significantly different stock levels than the rest of country X. This conclusion further proves that we can't rely on the nominal mean MOS value alone to understand the data.

ii. T-Test: Turgistan regional MOS versus country-wide MOS

Based on the nominal mean of Turgistan, we can set up the following hypothesis:

Null Hypothesis: The average MOS in Turgistan is not significantly different than the average MOS in country X.
Alt Hypothesis: The average MOS in Turgistan is significantly worse than the average MOS in country X.

It appears that when aggregating all products, Turgistan has on average about 21 less days of stock than the rest of the country. To determine significance in this difference of average MOS between Turgistan and all other regions, we set a level of significance at 5%.

#134	print(sp.stats.ttest_ind(molvania_mos_array, mos_array))

Our t-test returns statistic=-2.082 and p-value=0.037.
Since (p-value < 0.05), we reject the null hypothesis, and accept that on average, Turgistan has significantly lower stock levels than the rest of country X.

b. Regressions

i. Potential regression with georeferenced data

If this dataset were to have georeferenced data, we could potentially calculate the distances from facilitates to their supply points. It would be interesting to investigate if more remote facilities have worse AMI. We could conduct a linear regression to see if an increase in distance between supply points yields a worse AMI. This insight could help us understand if distribution operations should be improved in remoter areas.

ii. Time series analysis

Time series analysis would provide interesting insights on how monthly stock levels change over time, giving a general idea on whether stock management is improving or deteriorating over time in the country. However, the data we are working with here was collected in only one period, October 2018, so we are unable to conduct time series analysis on MOS data.

iii. Linear regression: does an increase in AMI yield better MOS?

With the data columns we have, there's not many interesting options for linear regressions. However, an option to conduct a regression would be to investigate the effect of AMI on MOS. Before looking at the data, it seems like common sense to assume that higher AMI would cause better MOS. But if the regression shows us that higher AMI does not result in higher MOS, that would be even more interesting because it would open-up opportunities for further analyses to understand why increased AMI does not yield better monthly stocks in country X.

#136-139	ami_array = np.array(df['AMI']) mos_array = np.array(df['MOS']) print(stats.linregress(ami_array, mos_array))

When we conduct the linear regression above, for all products in all facilities in all regions, we find that for everyone increase of AMI, MOS actually decreases ever so slightly by -0.008. While surprising, the regression results are not statistically significant at any level, so we conclude that an increase or decrease in AMI has a yet undetermined effect on average MOS.