Q: The average amount of rainfall in inches for each of 70 U.S. and Puerto Rico cities has been give:

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **7** | **7.2** | **7.8** | **7.8** | **11.5** | **13** | **14** | **14.6** | **15** |
| **30.2** | **30.6** | **30.8** | **31** | **31.4** | **31.7** | **32.5** | **33.4** | **34.4** |
| **38.8** | **38.9** | **39** | **39.9** | **40.2** | **40.2** | **40.8** | **40.8** | **41.8** |
| **48.3** | **48.5** | **49.1** | **49.2** | **54.5** | **54.7** | **56.8** | **59.2** | **59.8** |
| **17.2** | **17.4** | **20.7** | **22.9** | **24.7** | **25.9** | **29.1** | **30.2** | **48.2** |
| **35.9** | **35.9** | **36.1** | **36.2** | **36.2** | **37** | **37** | **37.6** | **38.7** |
| **42.6** | **42.7** | **42.8** | **43.1** | **43.4** | **44.7** | **45.5** | **46** | **46.4** |
| **15.2** | **35** | **42.5** | **67** | **16.2** | **35.1** | **42.5** |  |  |

1. Find the descriptive statistics for the data.

**Descriptive Statistics includes:**

* Mean
* Median
* Mode
* Variance
* Standard Deviation
* Skewness
* Kurtosis

1. **MEAN:**

The mean is used for several reasons:

* **Central Tendency**: It provides a single representative value that summarizes the entire dataset, indicating where most of the data points lie.
* **Easy Interpretation**: The mean is intuitive and easy to understand. It gives a quick idea of the typical value of the dataset.
* **Mathematical Properties**: Many statistical techniques and formulas are based on the mean. It has well-defined mathematical properties that make it useful for various calculations and analyses.
* **Comparison**: Mean allows for easy comparison between different datasets or different parts of the same dataset. It helps identify trends or differences.
* **Predictive Power**: In many cases, the mean can be used to predict future outcomes or estimate unknown values.

Overall, the mean is a fundamental statistical measure that is widely used in various fields, including science, economics, engineering, and social sciences, due to its simplicity, interpretability, and usefulness in **data analysis.**

1. **MEDIAN:**

* **Balanced Middle:** The median is a number that sits right in the middle of a list of numbers when they're arranged from smallest to largest.
* **Outlier Protection:** It doesn't get thrown off by really big or really small numbers in the list, so it's good for telling us what's typical even when there are some unusual values.
* **Fair Split:** Imagine if you had to split a group of people in half based on their heights. The median height would be the height of the person right in the middle, which seems fair, right?
* **Easy Comparison:** It's handy when comparing things like incomes or ages because it tells us what's typical without letting super high or super low values sway the result too much.

So, the median helps us find a fair middle point in a list of numbers, even if there are some really big or small numbers in there. It's useful for understanding what's typical or average in a dataset.

1. **MODE:**

* **Most Common Value**: The mode is the number that shows up the most times in a list of numbers.
* **Not Affected by Outliers**: Unlike the mean and median, the mode doesn't care about the actual values of the numbers; it just looks for the one that appears most often.
* **Can Have Multiple Modes**: A dataset can have one mode, more than one mode (bimodal, trimodal, etc.), or no mode at all if all values occur with the same frequency.
* **Useful for Categorical Data**: It's particularly useful for categorical data, like colors or types of fruit, where you want to know which option is the most popular.

So, the mode helps us identify the most common value in a dataset, making it useful for understanding which values occur most frequently.

1. **VARIANCE:**

Variance is a measure of how much the values in a dataset vary or spread out from the mean.

* **Spread of Data**: Variance tells us how much the numbers in a dataset differ from the average (mean) value.
* **Squared Differences**: To calculate variance, we find the difference between each number and the mean, square those differences, and then average those squared differences.
* **Higher Variance, More Spread Out**: A high variance means that the numbers in the dataset are spread out widely from the mean, while a low variance means they are closer to the mean.
* **Units Squared**: Since variance involves squaring differences, its units are squared units of the original data. For example, if the data represents lengths in meters, the variance would be in square meters.

In summary, variance helps us understand how much the numbers in a dataset vary from the mean, providing insights into the spread or dispersion of the data.

1. **STANDARD DEVIATION:**

The square root of variance gives us the standard deviation, which is another measure of the spread of data but is in the same units as the original data.

1. **SKEWNESS:**

Skewness is a statistical measure that quantifies the asymmetry of the probability distribution of a real-valued random variable about its mean. In simpler terms, it indicates whether the data is skewed to the left or right of the mean, or if it's symmetric.

* If the distribution is skewed to the left (negatively skewed), it means that the left tail is longer or fatter than the right tail. The mass of the distribution is concentrated on the right-hand side.
* If the distribution is skewed to the right (positively skewed), it means that the right tail is longer or fatter than the left tail. The mass of the distribution is concentrated on the left-hand side.
* If the distribution is symmetric, it means that the left and right tails are balanced and have roughly the same shape.

1. **KURTOSIS:**

Kurtosis is another statistical measure used to describe the shape of a probability distribution. It quantifies the "peakedness" or "tailedness" of a distribution relative to the normal distribution.

* Positive kurtosis (leptokurtic): A distribution with positive kurtosis has a higher peak and heavier tails than the normal distribution. This means it has more extreme values in the tails.
* Negative kurtosis (platykurtic): A distribution with negative kurtosis has a flatter peak and lighter tails than the normal distribution. This indicates fewer extreme values in the tails.
* Mesokurtic: A distribution with zero kurtosis (or close to zero) is called mesokurtic and resembles a normal distribution in terms of its shape.

Top of Form

**COMMANDS IN PYTHON:**

import pandas as pd

import matplotlib.pyplot as plt

location = "C:\\Users\HP\Documents\mybook.xlsx"

df = pd.read\_excel(location)

column\_data = df["Amount of Rainfall"]

mean = column\_data.mean()

median = column\_data.median()

mode = column\_data.mode()[0]

# mode() returns a Series, get the first mode if there are multiple

std\_dev = column\_data.std()

variance = column\_data.var()

print(f"The value of mean is : {mean}")

print(f"The value of median is : {median}")

print(f"The value of standard deviation is: {std\_dev}")

print(f"The value of variance is : {variance}")

print(f"The value of mode is : {mode}")

# Create a figure and a set of subplots

fig, axs = plt.subplots(nrows=1, ncols=2, figsize=(10, 5))

# Adjust figsize to fit your screen

# Plotting the histogram on the first subplot

axs[0].hist(column\_data, bins=20, alpha=0.7, color='blue', edgecolor='black')

axs[0].set\_title('Distribution of Amount of Rainfall')

axs[0].set\_xlabel('Amount of Rainfall')

axs[0].set\_ylabel('Frequency')

axs[0].axvline(mean, color='red', linestyle='dashed', linewidth=1, label=f'Mean: {mean:.2f}')

axs[0].axvline(median, color='green', linestyle='dashed', linewidth=1, label=f'Median: {median:.2f}')

axs[0].axvline(mode, color='orange', linestyle='dashed', linewidth=1, label=f'Mode: {mode:.2f}')

axs[0].legend()

# Creating a box plot on the second subplot

axs[1].boxplot(column\_data, vert=True, patch\_artist=True)

axs[1].set\_title('Box Plot of Amount of Rainfall')

axs[1].set\_ylabel('Amount of Rainfall')

axs[1].set\_xticks([1])

axs[1].set\_xticklabels(['Rainfall'])

axs[1].grid(True)

plt.tight\_layout() # Adjust the layout to make sure everything fits without overlapping

plt.show() # Display the plots

1. Describe the nature of the data using descriptive statistics, skewness and kurtosis.

* In these results, the mean is 34.89 and the median is 36.6. As the median is greater than the mean the data appears to be skewed to the left.
* The standard deviation is 13.71. With normal data, most of the observations are spread within 6 standard deviations on each side of the mean.
* As the skewness is -0.3, data is negatively skewed.
* As the kurtosis is -0.24, it is playtikurtic distribution (less peaked).

1. Plot histogram, Box plot.

**Histogram of X Boxplot of X**



