statistics for data analyst

why statistics is important ?

1. Customer Preferences

Meet Sarah, the manager of a local coffee shop called "Cozy Beans." Sarah has noticed that some customers visit her shop every day, while others only stop by occasionally. She wants to increase customer loyalty and encourage more frequent visits, so she decides to create a new loyalty program.

She isn't sure what rewards or benefits would be most appealing to her customers. Should she offer free drinks after a certain number of purchases? Or maybe discounts on pastries or merchandise?

Without understanding her customers' preferences,

Sarah risks investing time and resources into a loyalty program that may not attract much interest or participation.

To solve this problem, Sarah decides to gather data from her customers. She sets up a simple survey at the checkout counter, asking customers about their preferences for loyalty program rewards. She also tracks purchase behavior to see which items are most popular among different customer groups.

After collecting and analyzing the data using basic statistical

(data distribution - covariance & correlation) methods,

Sarah discovers that the majority of her customers are most interested in earning points towards free drinks rather than discounts on other items. Armed with this insight, Sarah designs her loyalty program to offer a free drink for every ten purchases.

Sarah was able to understand her customers' preferences and design an effective loyalty program that not only attracts new customers but also keeps existing ones coming back

2. ma7el 2asab

Sure, here's a real story:

In Egypt, "Juicy Delight," wanted to expand its product line to include a new flavor of juice. However, the company was unsure whether there would be enough demand for the new flavor to justify the production costs.

To make an informed decision, Juicy Delight decided to use statistical analysis of sales data. They first conducted market research to identify potential flavors that would appeal to their target demographic. Based on the research findings, they narrowed down their options to three flavors: mango, guava, and pineapple.

Next, Juicy Delight launched a limited-time promotion where customers could sample all three flavors for free and provide feedback. They tracked the sales and collected feedback from customers at various locations across Egypt.

Using statistical analysis, Juicy Delight analyzed the sales data to determine which flavor was the most popular among customers. They calculated metrics such as total sales volume, average sales per location, and customer preferences by # but there are problem , they found it depend on location of their branch (tagamo3 highest paid --> pistachio while masr el jadida --> الفستق)

Juicy Delight could use the topic of covariance & correlation to analyze the relationship between different variables related to their sales data.

For example, Juicy Delight could explore the correlation between factors such as promotional activities (e.g., offering free samples), geographical locatioand sales volume of different flavors. By calculating the covariance and correlation coefficients between these variables, Juicy Delight can determine if there are any significant relationships or patterns that can

help them optimize their marketing strategies and product offerings.

what is statistics

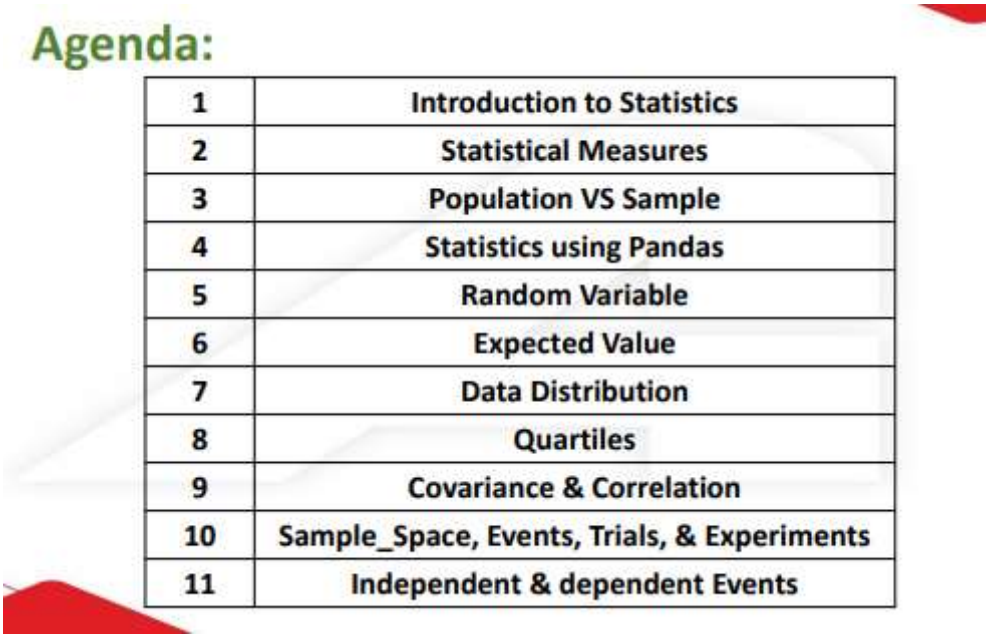
study of collecting, analyzing, interpreting (تفسير), and presenting data to make informed decisions.

another defination

Statistics is the science of summarizing and describing the data.

why we use statistics in data analyst?

to understand data better and make smarter decisions.

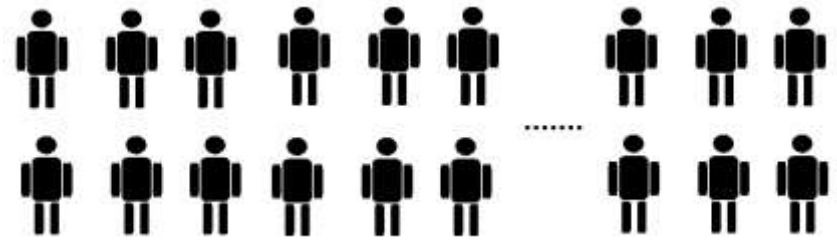


statistical measure

A Statistical Measure is a value, that is calculated to summarize many records(rows) of information into one single valu

example

Suppose you have a dataset that contains about 100,000,000 observations about Egyptian people height



If you want to describe what is the height of Egyptian people are, you don't tell the height of each single person of the 100,000,000 people in the Egyptian population! But instead, you simply say "The average height of the Egyptian people is 170cm".

What you have just done is that you summarized the 100,000,000 observations into one number, 170cm, which we call a statistical measure.

is that mean that any summarized data (statistical measure) always must be number like 170 cm ?

big no

statistical measure can have diffrent shape of data according to type of data

there are two main types of data

1. Continous data (numerical)

2. discrete data (categorical)

Continuous Data	Discrete Data
➢ Is the data that has infinite number of possible values. ➢ Also known as **Numerical data**. ➢ Continuous data could be: ➢ Float dtypes; such as, Salary or Weight. ➢ Int dtypes that have large number of possible unique values; such as, number-of-hours-played.	➢ Is the data that has finite number of possible values ➢ Also known as **Categorical data**. ➢ Continuous data could be: ➢ String dtypes; such as, City-name. ➢ Int dtypes that have small number of possible unique values; such as, number-of-children.

okay so its name is statistical (((((measures))))) so

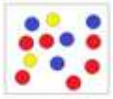
Popular Statistical Measures

1. Probability

2. Measures of Central Tendency.

3. Measures of dispersion (Deviation).

probability

Probability	Definition: Is the ratio between frequency of the unique-value & total number of samples **Example 1** Example1, suppose you have a dice:  ➢ The unique possible values are; 1, 2, 3, 4, 5, 6. ➢ Probability of 1 = 1 / 6 = .167 **Example 2** suppose you have the box of balls on the right:  ➢ The unique possible values are; blue, red, yellow. ➢ Probability of blue = 4 / 12 = .333

Measures of Central Tendency.

Measures of Central Tendency	Definition: Is the ratio between frequency of the unique-value & total number of samples **There are three main measures of central tendency:** ➢ Mean. (used to summarize numeric data) ➢ Median. (used to summarize numeric data) ➢ Model. (used to summarize categorical data) **Mean** **Definition:** ratio between the summation of all values and total number of observation in the data Example 1: suppose you have the following set of observation [5, 2, 3, 10, 20] Mean = (5+2+3+10+20) / 5 = 8. Mean is used with numerical data that doesn't contain extreme values (outliers), because mean is sensitive to outliers We use symbol μ to represent the mean.

Median

Definition: Median is the middle value in the data **after being sorted.**

Steps:

- First sort the data
- then Find the number in the middle
- this is your Median.

If there are two number in the middle, then the Median is the average between them

- Median is used with numerical data that contains outliers.

Example1	Example2
➢ Suppose you have this set of observations: [5, 2, 3, 10, 20] . ➢ First sort them ➔ [2, 3, 5, 10, 20]. ➢ Median = 5.	➢ Suppose you have this set of observations: [3, 5, 2, 3, 10, 20] . ➢ First sort them ➔ [2 , 3, 3, 5, 10, 20]. ➢ Median = (3+5) / 2 = 4.

Mode

- Mode is the most frequent value in the data.
- Mode is used with categorical data.

Example1	Example2
➢ Suppose you have this set of observations: [5, 2, 3, 3, 2, 3, 1, 5, 9, 8, 3, 1, 7, 6]. ➢ Mode= 5.	➢ Suppose you have this set of observations: ["Cairo", "Alex", "Aswan", "Alex", "Alex", "Mansoura", "Alex", "Cairo"]. ➢ Mode = "Alex".

measure of dispersion

Measures of dispersion (Deviation).

Definition:

- Are measures used to measure the spread of the data.
- Also Called Measures of Deviation.

For example, suppose you have the following two sets of numbers:

- Set1 = [5, 5, 5, 5, 5] & Set2 = [-5, 0, 5, 10, 15].
- The two sets contains the same value of mean = 5.
- But as you can see Set2 has more spread than set1.
- So, we need a way to measure the amount of spread.

There are two main measures of Dispersion:

- Variance.
- Standard deviation.

Standard Deviation is the most used as a measure of dispersion, that's why we call it standard, however variance is a popular measure too and has its applications

Variance

Definition: Is the average of all differences between each value in the data & the mean of this data

$\sigma 2$ is used to represent the Variance.

Formula: $\sigma^2 = \frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}$, where X_i represents the i^{th} value in the data, and N represents total number of values.

Example1	Example2
➢ Data = [5, 5, 5, 5, 5] .	➢ Data = [-5, 0, 5, 10, 15].
➢ μ = (5+5+5+5+5) / 5 = 5.	➢ μ = (-5+0+5+10+15) / 5 = 5.
➢ σ2 = ((5-5)² + (5-5)² + (5-5)² + (5-5)² + (5-5)²) / 5 = 0.	➢ σ2 = ((5--5)² + (5-0)² + (5-5)² + (5-10)² + (5-15)²) / 5 = 50.
➢ Variance = 0	➢ Variance = 50.

When variance equals zero, it means that there is no variability or spread in the data. In other words, all the values in the dataset are identical or constant.

When variance equals 50, it means that the numbers in the data are spread out from the average by an amount that, on average, equals 50 squared units.

Standard Deviation

➢ Is the square root of the variance

➢ σ is used to represent the Standard deviation.

Standard deviation is always preferred over variance as a measure of dispersion, and the reason is that unlike variance, standard deviation is not sensitive to outliers.

Formula: $\sigma = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}}$, where X_i represents the i^{th} value in the data, and N represents total number of values.

Example:

Measures of Dispersion (Standard Deviation):

Example1	Example2
➢ Data = [5, 5, 5, 5, 5] .	➢ Data = [-5, 0, 5, 10, 15].
➢ μ = (5+5+5+5+5) / 5 = 5.	➢ μ = (-5+0+5+10+15) / 5 = 5.
➢ σ2 = ((5-5)² + (5-5)² + (5-5)² + (5-5)² + (5-5)²) / 5 = 0.	➢ σ2 = ((5--5)² + (5-0)² + (5-5)² + (5-10)² + (5-15)²) / 5 = 50.
➢ σ = √σ2 = √0 = 0.	➢ σ = √σ2 = √50 = 7.07
➢ Standard deviation = 0.	➢ Standard deviation = 7.07

Standard deviation = 0 there is no variability or spread around the mean

Standard deviation = 7.07

Standard deviation = 7.07 mean the data points are about 7.07 units away from the mean.

population vs sample

Population	is the whole complete set of observation. **Example** ➢ In Egypt, we have 100,000,000 people if we could collect

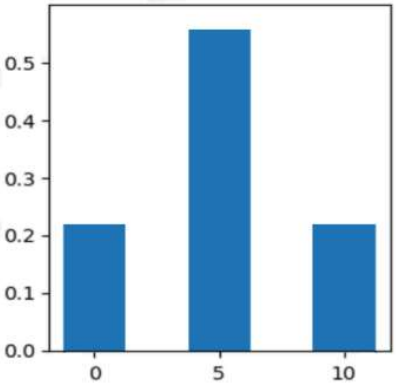
Expected Value Example:

- Suppose the following Random Variable:
  - X = [0, 5, 5, 5, 10, 0, 5, 10, 5].
  - P(X=0) = 2/9 = .222
  - P(X=5) = 5/9 = .556
  - P(X=10) = 2/9 = .222
  - E(X) = 0 * .222 + 5 * .556 + 10 * .222 = 5.

Data Distribution

Data Distribution is a way to describes how the observations are distributed or spread across the unique values of the data. ➢ In other words, Data Distribution represents how much each unique value occurs in the data or how frequent each unique value is.

- If you have Random Variable X = [0, 5, 5, 5, 10, 0, 5, 10, 5].
- Then the data distribution of this random variable is distributed as following:
  - 22.2% of the data belong to (X=0).
  - 55.6% of the data belong to (X=5).
  - 22.2% of the data belong to (X=0).



note

1. It's common to represent the data distribution as a graph called Histogram.
2. A histogram is a 2-dimensional graph, where: i. X-axis represents the unique values in the Random Variable. ii. Y-axis represents the probability of each unique value.

Data Distribution Types

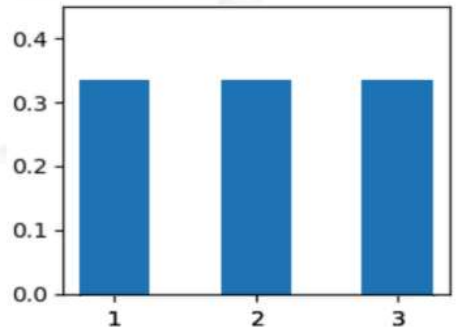
1. Uniform Distribution.
2. Normal Distribution.
3. Right-Skewed Distribution.
4. Left-Skewed Distribution.

1. Uniform Distribution.

observations are equally distributed among the unique values. In other words, all the unique values occur equally with the same frequency.

For example, Suppose you have X = [1, 2, 2, 3, 1, 3],
then the distribution is:

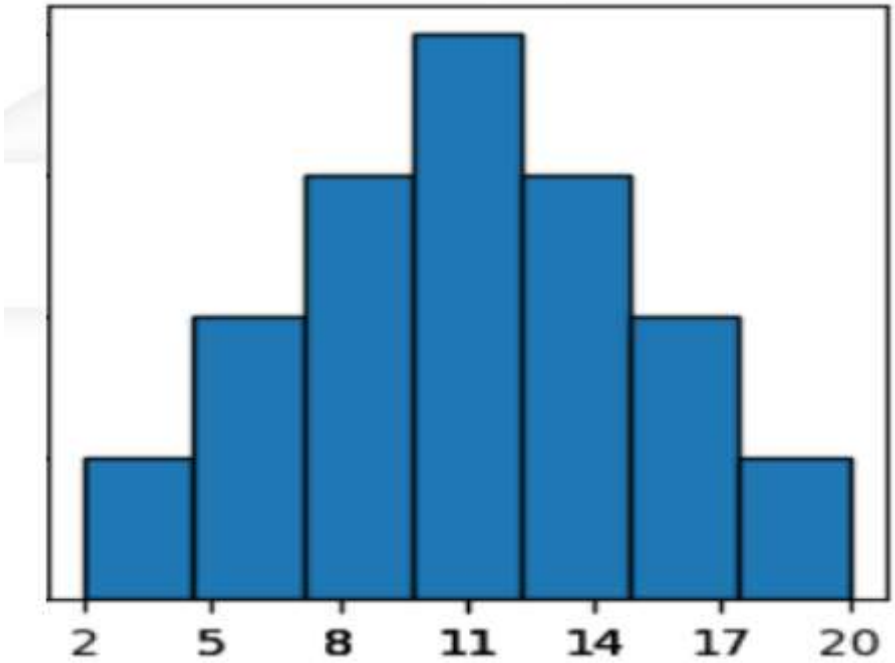
- 33.3% of the data belong to (X=1).
- 33.3% of the data belong to (X=2).
- 33.3% of the data belong to (X=3).



You can replace missing values with the mean or median.

2. Normal Distribution

Is Data Distribution where observations are distributed around the mean the most

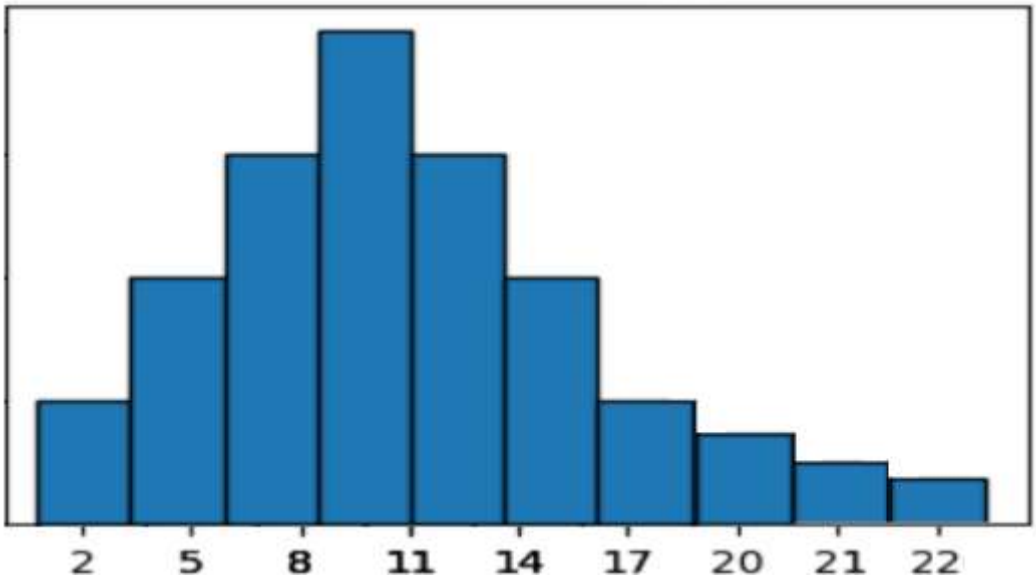


The distribution histogram takes a shape of symmetric bell.

you can also use the median, especially if there are outliers that might affect the mean significantly.

3. Left-Skewed Distribution:

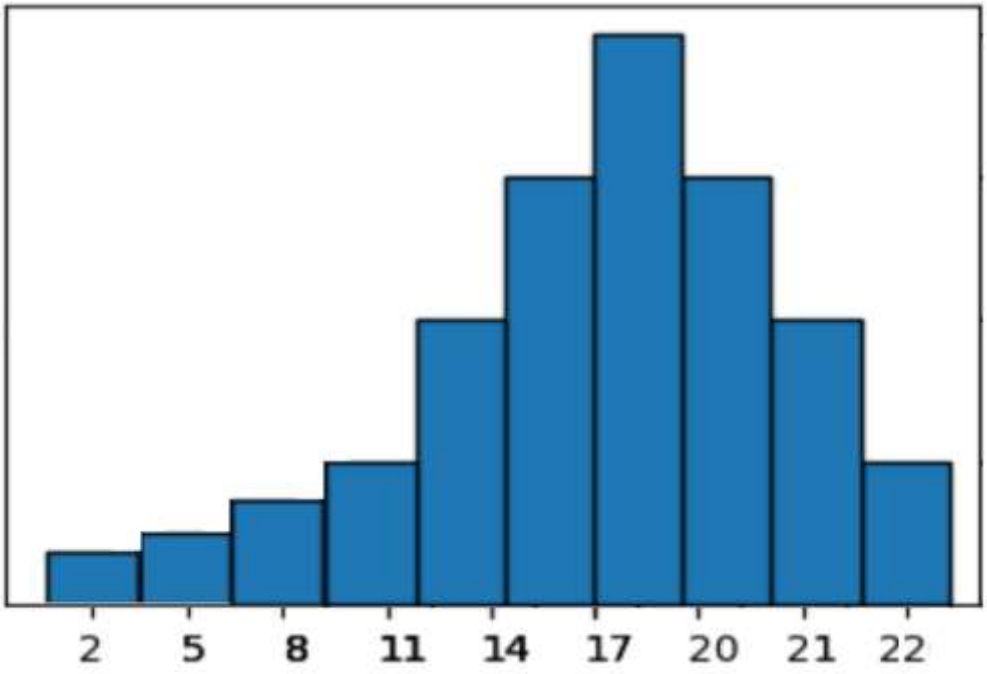
Data Distribution where observation are mostly distributed around mean and left side to the mean, with few observations at the extreme right to the mean



Replace missing values with the median, as it's less influenced by outliers present in the long tail.

4. Right-Skewed Distribution:

Is Data Distribution where observation are mostly distributed around mean and right side to the mean, with few observations at the extreme left to the mean.



Replace missing values with the median, as it's less influenced by outliers present in the long tail.

replacing missing data with mode when you deal with categoriacal data

outliers

Outliers are data points that are significantly different from other observations in a dataset, like a really tall person among a group of average-height people.

another defination

are extreme values that occur in the data.

example

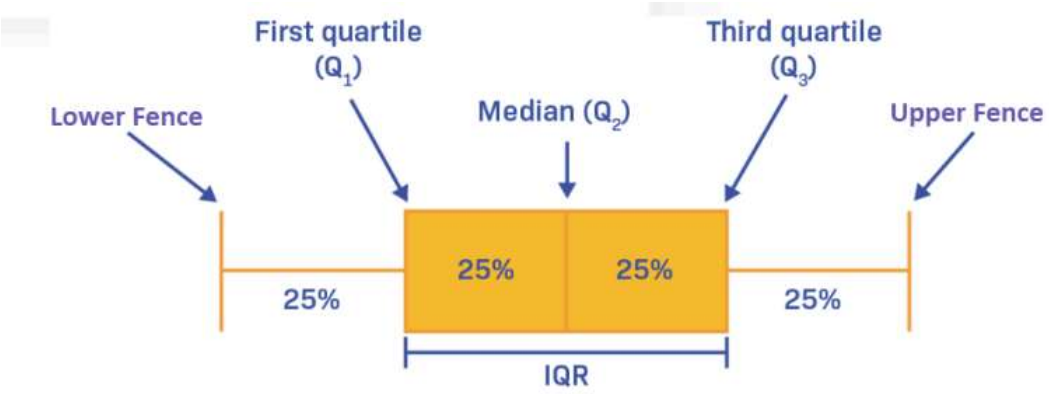
Suppose you have a random variable X=[20, 30, 10, 50, 180] where X represents people ages The value 180 is an outlier because it's a strange or extreme value, since it's no common to see a 180 years-old person

Quartiles

Quartiles are numbers used to detect fences or thresholds where if a number exceeds these fences or thresholds, then this number is considered to be an outlier

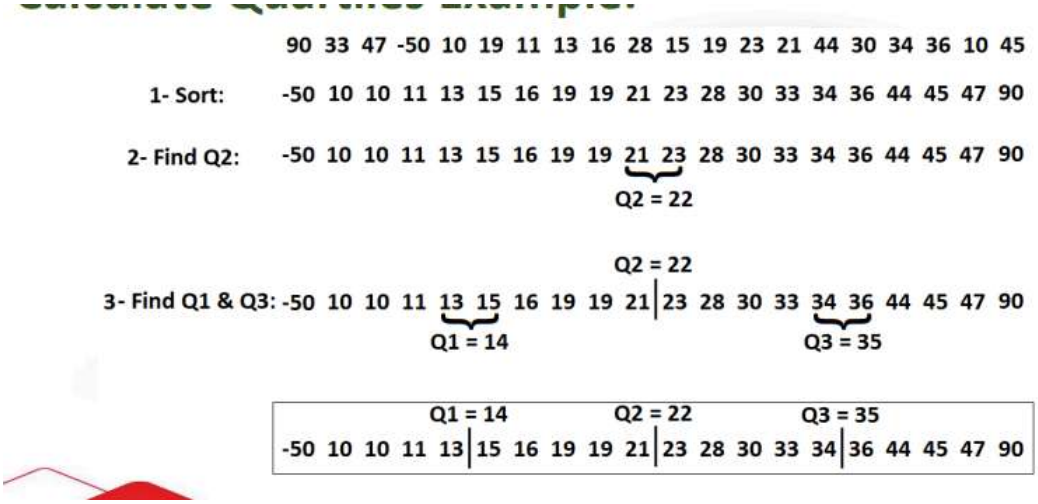
There are three types of quartiles to calculate to be able to calculate the fences :

1. First Quartile (Q1).
2. Second Quartile (Q2).
3. Third Quartile (Q3)



How to Calculate Quartiles?

1. Sort the Random Variable data (column).
2. Calculate the median of the Random Variable, and this is your Q2.
3. Calculate the median of the subset right to Q2, and this is your Q1.
4. Calculate the median of the subset left to Q2, and this is your Q3



Outlier fences

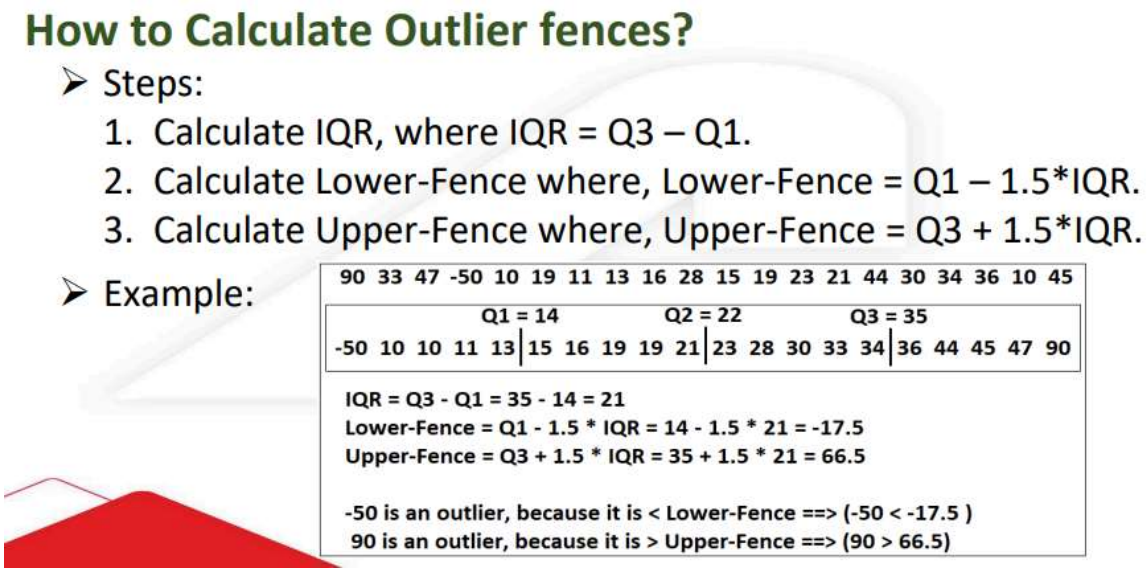
two fences we need to calculate so that if a number exceed these fences, then it is considered an outlier.

1. Upper Fence:

If a number is larger than the upper fence, then it is considered an outlier.

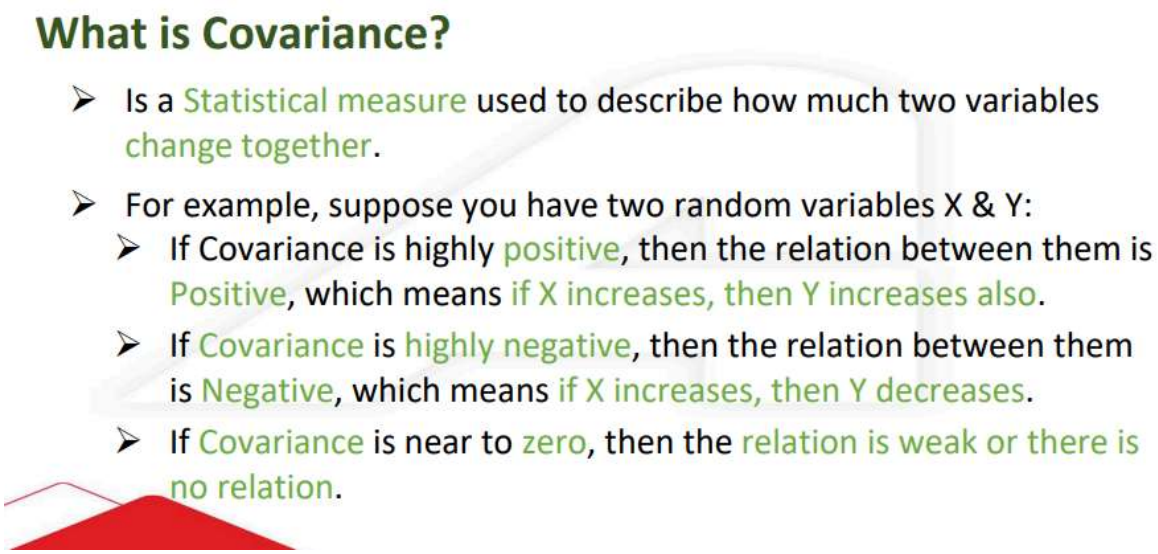
2. Lower Fence:

If a number is smaller than the lower fence, then it is considered an outlier



Covariance & Correlation

What is Covariance?



example

Engineering and Quality Control: Covariance is employed in quality control processes to analyze the relationship between different factors affecting product quality. For example, in manufacturing, covariance can help determine how changes in one process variable affect the quality of the final product.

How to Calculate Covariance?

- Formula:
  - $Cov(X, Y) = \sum_{i=1}^{n}((X_i - \mu_x) * (Y_i - \mu_y)) / n$.
  - n is the number of samples.
  - μ_x is the mean of Random Variable X.
  - μ_y is the mean of Random Variable Y.
- Example:

X = [1, 2, 3, 4, 5, 6, 7, 8, 9] Y = [9, 8, 7, 6, 5, 4, 3, 2, 1]
$\mu_x = 5$ $\mu_y = 5$ n = 9

Cov(X, Y) = ((1-5)*(9-5) + (2-5)*(8-5) + (3-5)*(7-5) + (4-5)*(6-5) + (5-5)*(5-5)
+ (6-5)*(4-5) + (7-5)*(3-5) + (8-5)*(2-5) + (9-5)*(1-5) +)/n
= -6.667

Result: Cov(X, Y) = -6.667 < 0.
Conclusion: The relation between X & Y is Negative.

What is Correlation?

What is Correlation?

- Is a Statistical measure that is the same as Covariance, except that Correlation is normalized, which give us sense about the relation strength.
- Normalized means that Correlation has values in range = [-1:1].
- For example, suppose you have two random variables X & Y:
  - If Correlation is near to 1, then the relation between them is Strong Positive. While If Correlation is near to 2, then the relation between them is Strong Negative.
  - If Correlation is near to 0, then the relation is weak.

diffrent between Correlation Vs Covariance:

Correlation Vs Covariance:

- Correlation has values in range [-1 : 1]. While Covariance had values between [∞, -∞].
- Having a range between -1 & 1 is very useful since this helps us know how much strong is the relation between the two variables.
- This is useful if I want to compare two relations. While in covariance this is not possible.
- Example:

Correlation	Covariance
Relation1 = .5 Relation2 = .25 Relation1 is twice strong as Relation2.	Relation1 = 5 Relation2 = 2.5 You can't tell how much Relation1 is stronger than Relation2.

how to calcualte correlation

How to Calculate Correlation?

- Formula:
  - Corr(X, Y) = Cov(X, Y)/(σ_x*σ_y).
  - σ_x is the Standard-deviation of Random Variable X.
  - σ_y is the Standard-deviation of Random Variable Y.
- Example:

X = [1, 2, 3, 4, 5, 6, 7, 8, 9] Y = [9, 8, 7, 6, 5, 4, 3, 2, 1]

σ_X = 2.582 σ_Y = 2.582

Cov(X, Y) = -6.667

Corr(X, Y) = Cov(X, Y) / (σ_X * σ_Y) = -6.667 / (2.582 * 2.582) = -1

Result: Corr(X, Y) = -1.

Conclusion: The relation between X & Y is Negative.

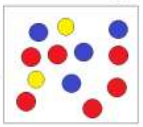
how to calcualte standard deviasion

Calculation	Formula	Notes
Population Standard Deviation	$\sigma = \sqrt{\frac{\Sigma (X_i - \mu)^2}{N}}$	μ = population average X = individual values in population N = count of values in population
Sample Variance	$s^2 = \frac{\Sigma (x_i - \bar{x})^2}{(n - 1)}$	$\bar{x}$ = sample average x = individual values in sample n = count of individual values in sample
Sample Standard Deviation	$s = \sqrt{\frac{\Sigma (x_i - \bar{x})^2}{(n - 1)}}$	$\bar{x}$ = sample average x = individual values in sample n = count of individual values in sample

Sample_Space, Events, Trials, & Experiments

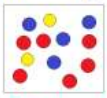
What is Sample Space?

- Is a set of all possible unique values of a Random Variable.
- We represent the sample space using S.
- Examples:

Example1	Example2
➢ suppose you are rolling a six-sided die. ➢ S = [1, 2, 3, 4, 5, 6]. 	➢ Suppose you have the following box of balls. ➢ S = [red, blue, yellow]. 

What are Events?

- An event is a subset of the sample space S.

Example1	Example2
➢ suppose you are rolling a six-sided die. ➢ S = [1, 2, 3, 4, 5, 6].  ➢ The possible events are: E1={1}, E2={2}, E3={3}, E4={4}, E5={5}, E6={6}, E7={1, 2}, …,E11={1, 3, 5}, etc. ➢ P(E7) means probability that die roll is 1 or 2. ➢ P(E11) means probability that die roll is an odd number.	➢ Suppose you have the following box of balls. ➢ S = [red, blue, yellow].  ➢ The possible events are: E1={red}, E2={red}, E3={red}, E4={red, blue}, E5={red, yellow}, E6={blue, yellow}, and E7={red, yellow, blue}. ➢ P(E6) means probability that you draw a blue ball or a yellow ball.

What are Trials?

- A trial is the act or the process we are doing, for example:
  - Flipping a coin is a trial.
  - Rolling a dice is a trial.
- The result of a trial is an event.
- For example, Suppose that a dice is rolled, and 5 appears:
  - Sample-Space = {1, 2, 3, 4, 5, 6}.
  - Trial = rolling the dice.
  - Event = {5}.

diffrent between independend events and dependendt event

Independent Events:

data analysis using statistics

udemy dataset

In [31]:
```
import pandas as pd
import numpy as np

df = pd.read_csv('Salaries.csv')
df.head()
```

```
C:\Users\Sameh Albadry\AppData\Local\Temp\ipykernel_15468\3726406145.py:4:
DtypeWarning: Columns (3,4,5,6,12) have mixed types. Specify dtype option o
n import or set low_memory=False.
  df = pd.read_csv('Salaries.csv')
```

Out[31]:

	Id	EmployeeName	JobTitle	BasePay	OvertimePay	OtherPay	Benefits	TotalPa
0	1	NATHANIEL FORD	GENERAL MANAGER-METROPOLITAN TRANSIT AUTHORITY	167411.18	0.0	400184.25	NaN	567595.4
1	2	GARY JIMENEZ	CAPTAIN III (POLICE DEPARTMENT)	155966.02	245131.88	137811.38	NaN	538909.2
2	3	ALBERT PARDINI	CAPTAIN III (POLICE DEPARTMENT)	212739.13	106088.18	16452.6	NaN	335279.9
3	4	CHRISTOPHER CHONG	WIRE ROPE CABLE MAINTENANCE MECHANIC	77916.0	56120.71	198306.9	NaN	332343.6
4	5	PATRICK GARDNER	DEPUTY CHIEF OF DEPARTMENT, (FIRE DEPARTMENT)	134401.6	9737.0	182234.59	NaN	326373.1

In [32]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 148654 entries, 0 to 148653
Data columns (total 13 columns):
 #   Column            Non-Null Count   Dtype  
---  ------            --------------   -----  
 0   Id                148654 non-null  int64  
 1   EmployeeName      148654 non-null  object 
 2   JobTitle          148654 non-null  object 
 3   BasePay           148049 non-null  object 
 4   OvertimePay       148654 non-null  object 
 5   OtherPay          148654 non-null  object 
 6   Benefits          112495 non-null  object 
 7   TotalPay          148654 non-null  float64
 8   TotalPayBenefits  148654 non-null  float64
 9   Year              148654 non-null  int64  
 10  Notes             0 non-null       float64
 11  Agency            148654 non-null  object 
 12  Status            38119 non-null   object 
dtypes: float64(3), int64(2), object(8)
memory usage: 14.7+ MB
```

data preprocessing

datatypes

In [33]:
```
df["BasePay"] = pd.to_numeric(df["BasePay"],errors = "coerce")
df["OvertimePay"] = pd.to_numeric(df["OvertimePay"],errors = "coerce")
df["OtherPay"] = pd.to_numeric(df["OtherPay"],errors = "coerce")
df["Benefits"] = pd.to_numeric(df["Benefits"],errors = "coerce")
df.drop(['Notes'],axis = 1,inplace = True)

df.dtypes
```

Out[33]:
```
Id                    int64
EmployeeName         object
JobTitle             object
BasePay             float64
OvertimePay         float64
OtherPay            float64
Benefits            float64
TotalPay            float64
TotalPayBenefits    float64
Year                  int64
Agency               object
Status               object
dtype: object
```

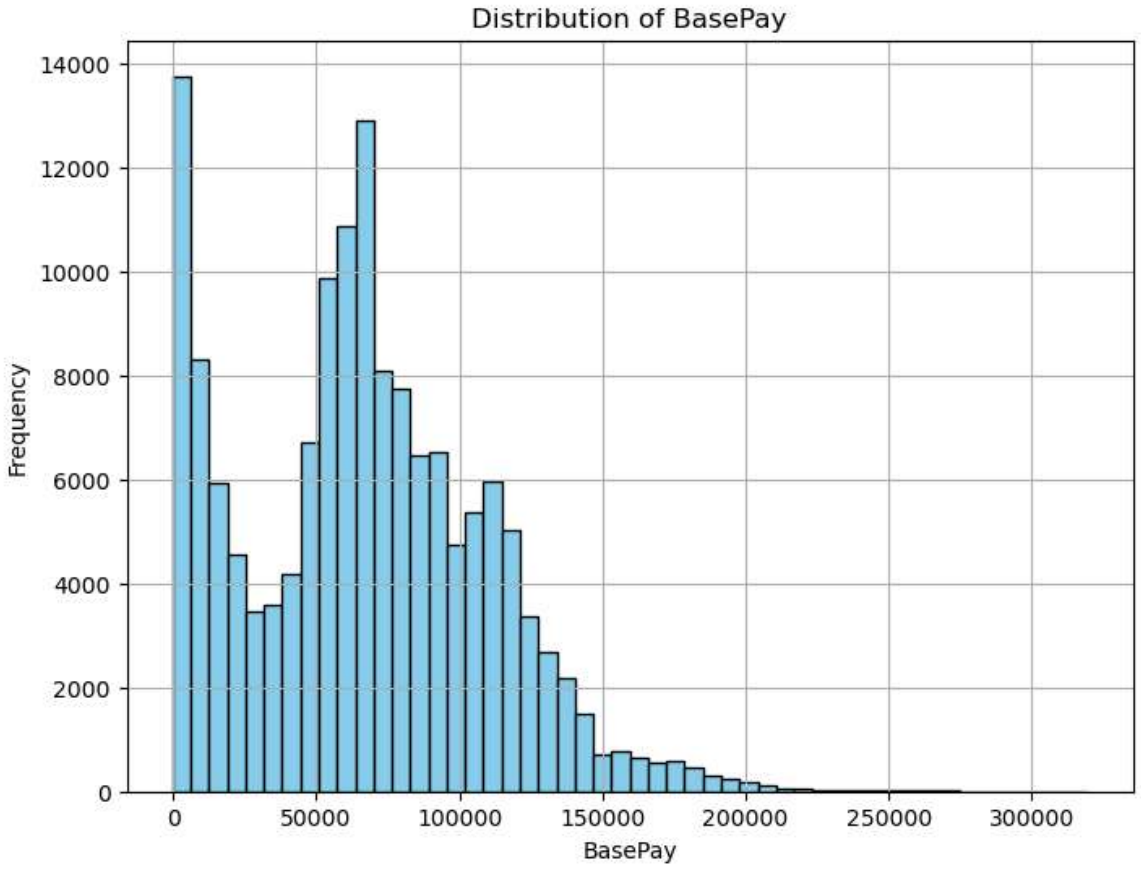
missing data

In [34]:
```
df.isnull().sum()
```

Out[34]:
```
Id                     0
EmployeeName           0
JobTitle               0
BasePay              609
OvertimePay            4
OtherPay               4
Benefits           36163
TotalPay               0
TotalPayBenefits       0
Year                   0
Agency                 0
Status            110535
dtype: int64
```

In [37]:
```
import pandas as pd
import matplotlib.pyplot as plt
# Load the dataframe

# Plot the distribution of the 'BasePay' column
plt.figure(figsize=(8, 6))
plt.hist(df['BasePay'], bins=50, color='skyblue', edgecolor='black')
plt.xlabel('BasePay')
plt.ylabel('Frequency')
plt.title('Distribution of BasePay')
plt.grid(True)
plt.show()
```



left skew distribution

therfore replace nan with median

In [35]:
```
median_basepay = df['BasePay'].median()
median_basepay
```

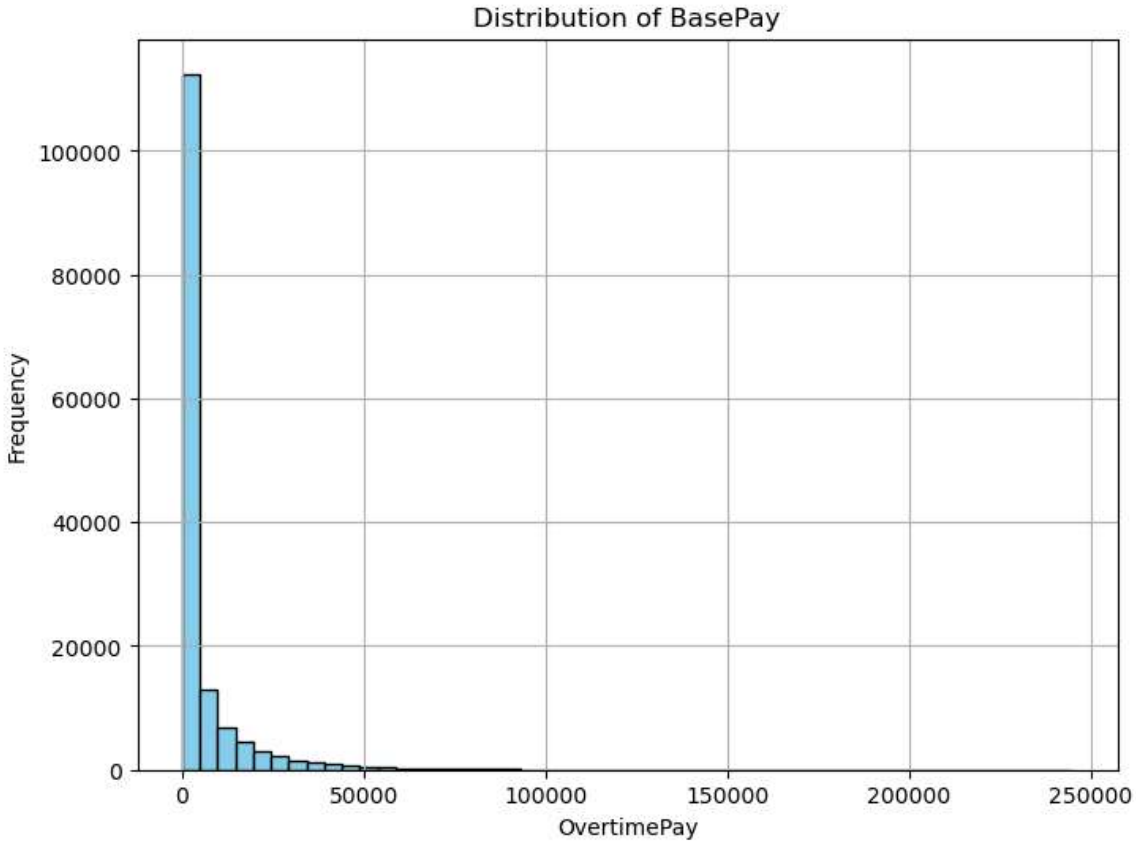
Out[35]:
```
65007.45
```

In [36]:
```
df['BasePay'].fillna(median_basepay, inplace=True)
df.isnull().sum()
```

Out[36]:
```
Id                      0
EmployeeName            0
JobTitle                0
BasePay                 0
OvertimePay             4
OtherPay                4
Benefits            36163
TotalPay                0
TotalPayBenefits        0
Year                    0
Agency                  0
Status             110535
dtype: int64
```

In [38]:
```
import pandas as pd
import matplotlib.pyplot as plt
# Load the dataframe

# Plot the distribution of the 'BasePay' column
plt.figure(figsize=(8, 6))
plt.hist(df['OvertimePay'], bins=50, color='skyblue', edgecolor='black')
plt.xlabel('OvertimePay')
plt.ylabel('Frequency')
plt.title('Distribution of BasePay')
plt.grid(True)
plt.show()
```

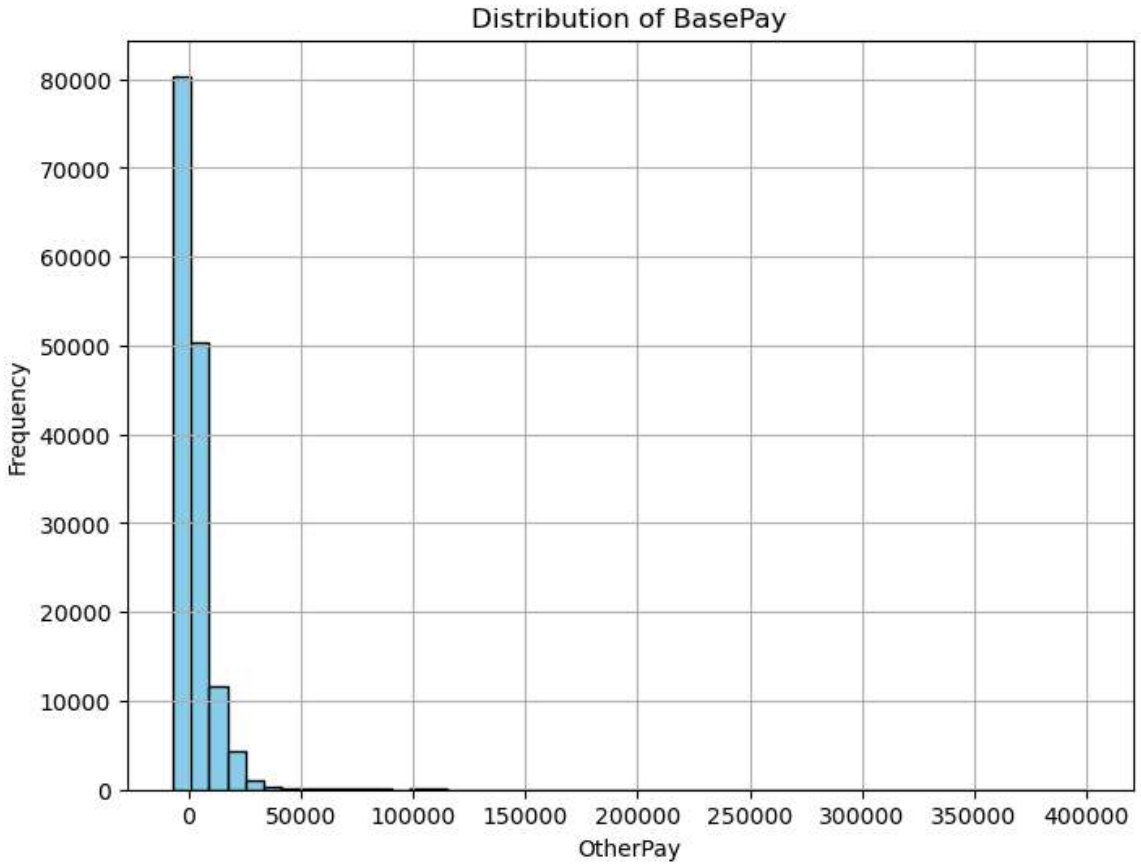


In [39]:
```
median_OvertimePay = df['OvertimePay'].median()
df['OvertimePay'].fillna(median_basepay, inplace=True)
df.isnull().sum()
```

Out[39]:
```
Id                      0
EmployeeName            0
JobTitle                0
BasePay                 0
OvertimePay             0
OtherPay                4
Benefits            36163
TotalPay                0
TotalPayBenefits        0
Year                    0
Agency                  0
Status             110535
dtype: int64
```

In [40]:
```
import pandas as pd
import matplotlib.pyplot as plt
# Load the dataframe

# Plot the distribution of the 'BasePay' column
plt.figure(figsize=(8, 6))
plt.hist(df['OtherPay'], bins=50, color='skyblue', edgecolor='black')
plt.xlabel('OtherPay')
plt.ylabel('Frequency')
plt.title('Distribution of BasePay')
plt.grid(True)
plt.show()
```

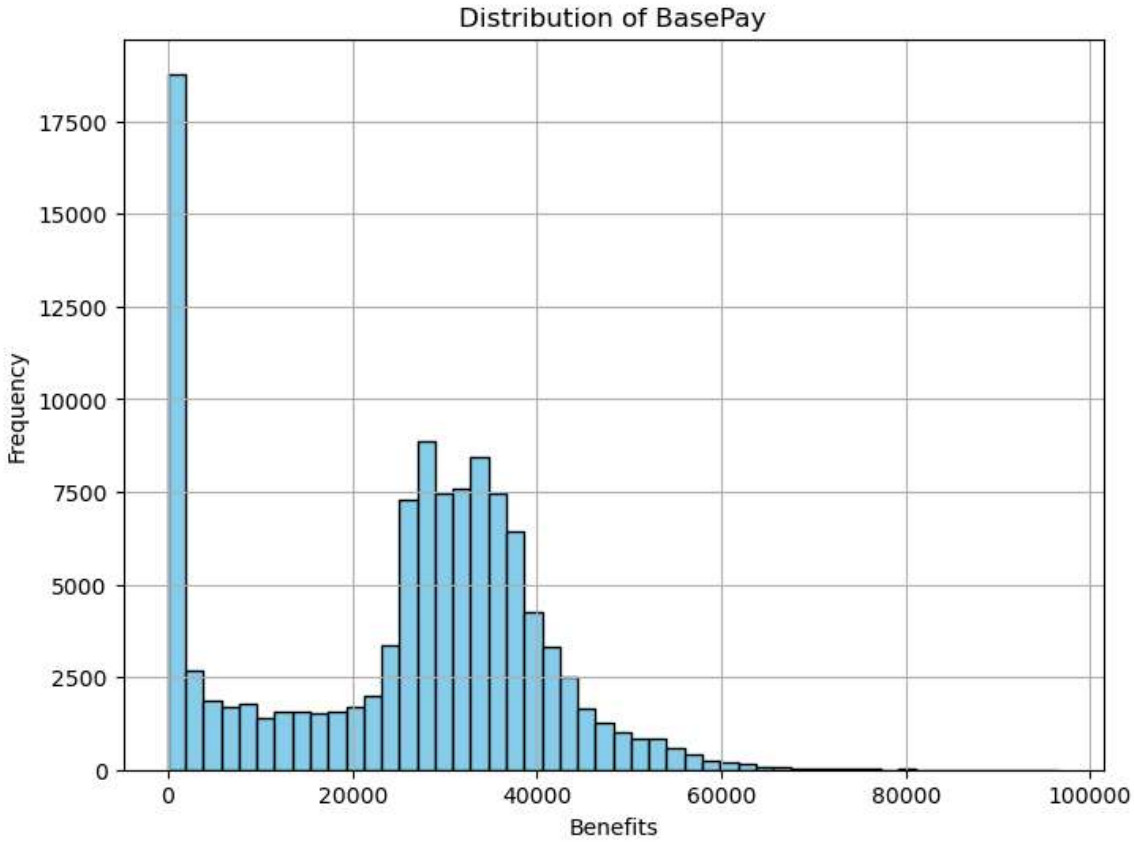


In [41]:
```
median_OvertimePay = df['OtherPay'].median()
df['OtherPay'].fillna(median_basepay, inplace=True)
df.isnull().sum()
```

Out[41]:
```
Id                     0
EmployeeName           0
JobTitle               0
BasePay                0
OvertimePay            0
OtherPay               0
Benefits           36163
TotalPay               0
TotalPayBenefits       0
Year                   0
Agency                 0
Status            110535
dtype: int64
```

In [42]:
```
import pandas as pd
import matplotlib.pyplot as plt
# Load the dataframe

# Plot the distribution of the 'BasePay' column
plt.figure(figsize=(8, 6))
plt.hist(df['Benefits'], bins=50, color='skyblue', edgecolor='black')
plt.xlabel('Benefits')
plt.ylabel('Frequency')
plt.title('Distribution of BasePay')
plt.grid(True)
plt.show()
```



In [43]:
```
median_OvertimePay = df['Benefits'].median()
df['Benefits'].fillna(median_basepay, inplace=True)
df.isnull().sum()
```

Out[43]:
```
Id                    0
EmployeeName          0
JobTitle              0
BasePay               0
OvertimePay           0
OtherPay              0
Benefits              0
TotalPay              0
TotalPayBenefits      0
Year                  0
Agency                0
Status           110535
dtype: int64
```

In [46]:
```
df['Status'].fillna(0,inplace = True)
df.isnull().sum()
```

Out[46]:
```
Id                 0
EmployeeName       0
JobTitle           0
BasePay            0
OvertimePay        0
OtherPay           0
Benefits           0
TotalPay           0
TotalPayBenefits   0
Year               0
Agency             0
Status             0
dtype: int64
```

In [ ]: