Report

1. Basic Data Exploration:

We started to explore the data and found that it has (148654, 13) size "148654 rows, 13 columns"

When we explored the type of each column we found the following:

- The columns (Id and Year) have type int64.
- The columns (BasePay, OvertimePay, OtherPay, Benefits, TotalPay, TotalPayBenefits, Notes, Status) have type float64.
- And other columns (EmployeeName, JobTitle, Agency) have type object.

To check if there are any missing values in each column:

We call the .info() function that in column "Non-Null" prints the numbers of non-null values in column

In another way "dfSalaries.isnull().sum()" returns only the number of null values in each column, and we found:

- Notes and Status are null entire
- Benefits have 36163 null values.
- BasePay has 609 null values.
- OvertimePay and OtherPay have 4 null values in their columns.

Visualize previous statistics using "msno.matrix(dfSalaries)" which plots the null value in the white line and the non-null in the black line.

2. Descriptive Statistics:

we called the describe function on the TotalPay column and got this result:

count	148654.000000
mean	74768.321972
Std	50517.005274
min	-618.130000
25%	36168.995000
50%	71426.610000
75%	105839.135000
max	567595.430000

3. Data cleaning:

to handle Null Values:

- We drop Notes and Status columns.
- We fill OtherPay, OvertimePay, BasePay, and Benefits in the mean value of their column.

4. Basic Data Visualization:

- We Create histograms or bar charts to visualize the distribution of salaries:
  Based on the plot the distribution is as close to the *lognormal distribution*

- We create pie charts to represent the proportion of employees in different departments:
  we found that the chart Is not understandable (because there are 2159 unique values in the JobTitle column)
- So we plot only the top 10

5. Grouped Analysis:

-We group the data by Year column:
summary statistics:

count	4.000000
mean	435418.700000
Std	105379.725493
min	339282.070000
25%	356954.012500
50%	417398.650000
75%	495863.337500
max	567595.430000

Distribution of BasePay is: *Uniform distribution*

-We group the data by JobTitle column:
summary statistics:

count	2159.000000
mean	105535.306596
Std	55287.382189
min	0.000000
25%	71982.145000
50%	94078.010000
75%	126233.580000
max	567595.430000

Distribution of BasePay is: *Skewed normal distribution*

6. Simple Correlation Analysis:

To Identify any correlation between salary and another numerical column:

We used "dfSalaries.corr()['TotalPay']" which returns a number (between -1 and 1) for each column with the TotalPay column, if the number above 0.7 that's means there is a strong correlation.

We found there is a strong correlation to TotalPay with (Benefits: 0.78) and (BasePay: 0.95) and (TotalPayBenefits: 0.97)

Then we plot a scatter plot to visualize the relationship:

From the plot, we determined that the Correlation is Positive (with Benefits and BasePay and TotalPayBenefits)

The End
