Employee Dataset Report.

- **Introduction:**

  CSV file was the data format that I deal with, which contains information about employees, their job titles, salaries, and benefit, with 148654 rows, 13 columns: 'Id', 'EmployeeName', 'JobTitle', 'BasePay', 'OvertimePay', 'OtherPay', 'Benefits', 'TotalPay', 'TotalPayBenefits', 'Year', 'Notes', 'Agency', 'Status'.

  By the following tasks I described the data statistically, cleaned the data (specifically for the columns that have missing values) , visualized the data, grouped the data by one column, and finally I found the relation between two columns.

- **Task 1:**

  Data Exploration: As a basic step to explore the dataset, and this step gives these results:

  - Number of (rows , columns) = (148654, 13).
  - Dataset information:

#	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	148045 non-null	float64
4	OvertimePay	148650 non-null	float64
5	OtherPay	148650 non-null	float64
6	Benefits	112491 non-null	float64
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	0 non-null	float64

dtypes: float64(8), int64(2), object(3)
memory usage: 14.7+ MB

- **Task 2:**

  Descriptive Statistics: In Descriptive statistics you are describing, presenting, summarizing, and organizing your data, either through numerical calculations or graphs or tables. It analysis helps us to understand our data and is a very important part of Machine Learning.
  by applied a block of codes that are in the file of codes I got this results:

	Id	BasePay	OvertimePay	OtherPay	Benefits	TotalPay	TotalPayBenefits	Year	Notes	Status
count	148654.000000	148045.000000	148650.000000	148650.000000	112491.000000	148654.000000	148654.000000	148654.000000	0.0	0.0
mean	74327.500000	66325.448840	5066.059886	3648.767297	25007.893151	74768.321972	93692.554811	2012.522643	NaN	NaN
std	42912.857795	42764.635495	11454.380559	8056.601866	15402.215858	50517.005274	62793.533483	1.117538	NaN	NaN
min	1.000000	-166.010000	-0.010000	-7058.590000	-33.890000	-618.130000	-618.130000	2011.000000	NaN	NaN
25%	37164.250000	33588.200000	0.000000	0.000000	11535.395000	36168.995000	44065.650000	2012.000000	NaN	NaN
50%	74327.500000	65007.450000	0.000000	811.270000	28628.620000	71426.610000	92404.090000	2013.000000	NaN	NaN
75%	111490.750000	94691.050000	4658.175000	4236.065000	35566.855000	105839.135000	132876.450000	2014.000000	NaN	NaN
max	148654.000000	319275.010000	245131.880000	400184.250000	96570.660000	567595.430000	567595.430000	2014.000000	NaN	NaN

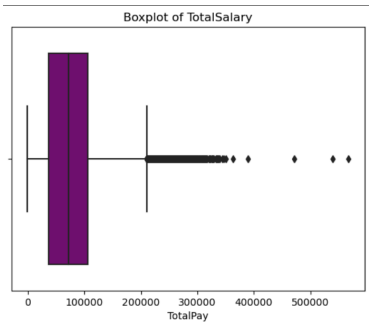
  For salary attribute:
  - Max: 567595.43
  - Min: -618.13
  - Central tendency:
  - Median: 71426.60999999999
  - Mean: 74768.32197169267
  - Mode: 0 0.0
  - Name: TotalPay, dtype: float64
  - Variance: 2551967821.8482866
  - std: 50517.005273949944
  - Percentiles [25th, 50th, 75th]:
  - [36168.995, 71426.61, 105839.135]
  - Salaries_range: 568213.56

- **Task 3:**
  Data Cleaning: it's a very important step to fix or remove incorrect, corrupted, incorrectly formatted, duplicate, or incomplete data within a dataset. The issue here is : Handle missing data in suitable method
  from (Task 1) I noticed that these columns have a null or missing values:
  BasePay, OvertimePay, OtherPay,Benefits.
  the common thing between these columns is that they all have the same numerical data type (float64), and that is mean I can solve this problem in different ways, such as:
  1. the attribute mode. 2. the attribute mean.
  and more, but because we have numerical values the second choice will be better, because (mean) minimizes the error in predicting the value in any dataset and the reason behind having the lowest error is that it includes every value in your data set as part of the calculation.
  As you can see in the following result, these columns (BasePay, OvertimePay, OtherPay,Benefits) have been filled:

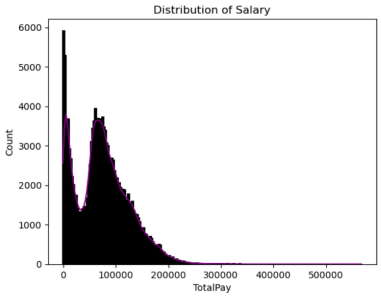
```
Data columns (total 13 columns):
 #   Column            Non-Null Count   Dtype
---  ------            --------------   -----
 0   Id                148654 non-null  int64
 1   EmployeeName      148654 non-null  object
 2   JobTitle          148654 non-null  object
 3   BasePay           148654 non-null  float64
 4   OvertimePay       148654 non-null  float64
 5   OtherPay          148654 non-null  float64
 6   Benefits          148654 non-null  float64
 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            0 non-null       float64
```

- **Task 4:**
  Basic Data Visualization: the graphical representation of data by using visual elements like charts, graphs, and maps. It is essential to analyze massive amounts of information and make data-driven decisions. One of the main goals of data visualization is communicating your results or findings with your audience.
  For salary attribute, there are a lot of ways to visualize it, such as:

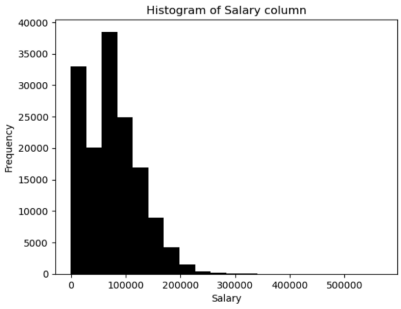
  1.Box plot



  2.Distribution plot



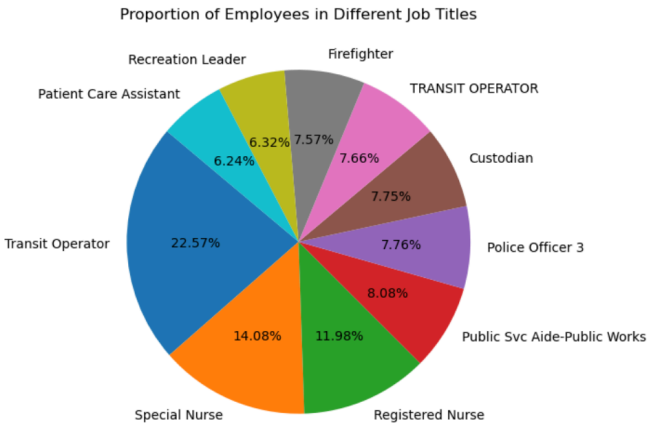
  3.histograms



The results of the visualization step, let us notice:

- There are a lot of outliers.
- The median of salaries approximately equal to 70000.
- The mean of salaries approximately equal to 75000.
- There are values very high like: 567595.43 (which will impact the mean, let it skewed, .. etc).
- The mean of salaries approximately equal to 75000.
- and more...

Then I generate a pie chart illustrating the proportion of employees in the top 10 job titles based on their frequency in the dataset.



This chart provides a visual representation of how employees are distributed across different job titles.

- **Task 5:**
  Grouped Analysis: to filter survey responses based on the type of response received on a particular question.
  In this task , I group the DataFrame by 'JobTitle' and calculate the mean 'TotalPay' for each group:

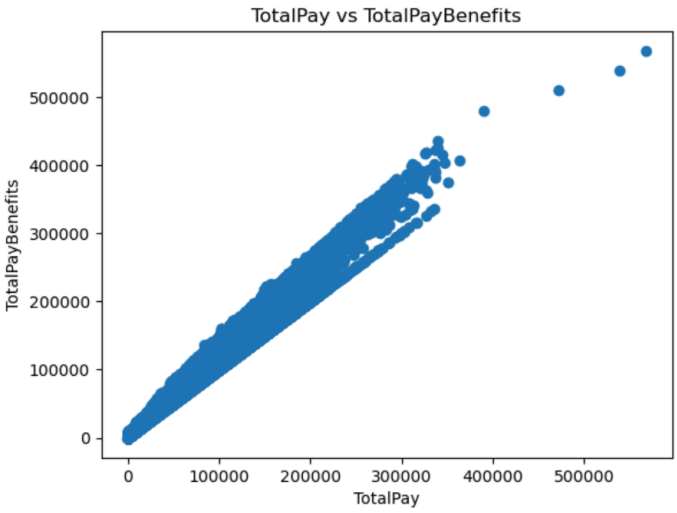
```
JobTitle
ACCOUNT CLERK                                    44035.664337
ACCOUNTANT                                       47429.268000
ACCOUNTANT INTERN                                29031.742917
ACPO,JuvP, Juv Prob (SFERS)                      62290.780000
ACUPUNCTURIST                                    67594.400000
                                                         ...
X-RAY LABORATORY AIDE                            52705.880385
X-Ray Laboratory Aide                            50823.942700
YOUTH COMMISSION ADVISOR, BOARD OF SUPERVISORS   53632.870000
Youth Comm Advisor                               41414.307500
ZOO CURATOR                                      66686.560000
Name: TotalPay, Length: 2159, dtype: float64
```

And here are summary statistics for each group, I take the JobTitle as a sample:

	Id								BasePay		Notes								Status			
	count	mean	std	min	25%	50%	75%	max	count	mean	...	75%	max	count	mean	std	min	25%	50%	75%	max
JobTitle																						
ACCOUNT CLERK	83.0	25734.819277	2621.592874	20766.0	24581.00	24774.0	26299.50	35638.0	83.0	43300.806506	...	NaN	NaN	0.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN
ACCOUNTANT	5.0	24159.200000	6787.702608	19264.0	19325.00	20993.0	25928.00	35286.0	5.0	46643.172000	...	NaN	NaN	0.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN
ACCOUNTANT INTERN	48.0	28128.833333	3941.782388	21536.0	23325.75	29979.5	31639.25	34267.0	48.0	28732.663958	...	NaN	NaN	0.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN
ACPO,JuvP, Juv Prob (SFERS)	1.0	135284.000000	NaN	135284.0	135284.00	135284.0	135284.00	135284.0	1.0	62290.780000	...	NaN	NaN	0.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN
ACUPUNCTURIST	1.0	18379.000000	NaN	18379.0	18379.00	18379.0	18379.00	18379.0	1.0	66374.400000	...	NaN	NaN	0.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
X-RAY LABORATORY AIDE	26.0	22502.076923	4694.321369	12940.0	19632.75	20927.0	24165.75	33182.0	26.0	47664.773077	...	NaN	NaN	0.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN
X-Ray Laboratory Aide	100.0	98890.620000	31404.565120	49777.0	65829.00	98379.5	130540.75	143632.0	100.0	46086.387100	...	NaN	NaN	0.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN
YOUTH COMMISSION ADVISOR, BOARD OF SUPERVISORS	1.0	23392.000000	NaN	23392.0	23392.00	23392.0	23392.00	23392.0	1.0	52609.910000	...	NaN	NaN	0.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Youth Comm Advisor	4.0	99677.750000	32595.496267	58843.0	88267.00	100704.5	112115.25	138459.0	4.0	39077.957500	...	NaN	NaN	0.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN
ZOO CURATOR	1.0	18779.000000	NaN	18779.0	18779.00	18779.0	18779.00	18779.0	1.0	43148.000000	...	NaN	NaN	0.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN

2159 rows × 80 columns

- **Task 6:**
  Simple Correlation Analysis: to measure how strongly two variables('TotalPay', 'TotalPayBenefits') are related,
  The correlation value = 0.9773128522072122
  Which means there is a very strong relation between them, and this scatter plot prof that:



DONE BY: SHAHED MOWAFFAQ ALZU'BI.