ShAI \_Assignment

salaries dataset

* Basic Data Exploration:

number of rows: 148654 and columns: 13

use df.info() to get more information about data (data type , miss value )

determine if that any duplication row and unique value

* Descriptive Statistics: use df.describe() to get all basic statistical information for numerical and textual data. deviation.
* Data Cleaning: 1) drop column:
  + - not have any data : [Status, Notes ]
    - all data have the same value : [Agency]
    - Does not provide important information for analysis ['EmployeeName' ,' Id']

2) drop row :

* Drop raw that have nan value “small row just 8 row from 14k

Drop null value in [OvertimePay , OtherPay]

3) full missing value using mathematical calculating :

* TotalPayBenefits = Benefits +TotalPay
* #TotalPay =BasePay + OvertimePay + OtherPay
* Fill miss value in [Benefits, BasePay ]
* Basic Data Visualization:
* Use histograms to visualize the distribution of salaries:

Because salary Continuous Data and histogram show the frequency

* pie charts to represent the employees in different departments:

first create new column for department extract from JobTitle

column. find that department name between () in JobTitel

* Simple Correlation Analysis: Create a heatmap of the correlation matrix

That show that are no relation between salary and years

use seaborn library to create scatter for the relationship

Between salary and other column

* Grouped Analysis:
  + - average total pay across different departments
    - average total pay across different departments across years
    - trend of average total pay across years
    - average benefits across different departments