بداية بعد قراءة البيانات قمنا باستعراض حجم البيانات وهي :

num of rows : 148654
num of columns : 13

بعد ذلك استعرضنا نوع البيانات في كل عمود وهي :

Id	int64
EmployeeName	object
JobTitle	object
BasePay	float64
OvertimePay	float64
OtherPay	float64
Benefits	float64
TotalPay	float64
TotalPayBenefits	float64
Year	int64
Notes	float64
Agency	object
Status	float64

من خلال استعراض عدد القيم None في كل عمود تبين ان كل من العمودين Notes و Status لا يحتويان أي قيم صحيحة لذلك يعتبر هذين العمودين بلا فائدة

ومن خلال استعراض عدد القيم المميزة في كل عمود نجد ان العمود Agency لا يحتوي سوى قيمة واحدة ثابتة وبذلك يعتبر عديم الفائدة

القيم الإحصائية إجمالي رواتب الموظفين :

mean : 74768.32197169267
median : 71426.60999999999
max : 567595.43
min : -618.13
standard deviation: 50517.005273949944

قمنا بحذف الاعمدة Agency و Status و Notes لأنها بلا فائدة كما بينا
في بقية العمدة تم استبدال كل القيم None بقيم صفرية لأن الاعمدة الاساسية التي تهم في الداتا لا تحوي أي قيم مفقودة لذلك من الخطأ حذف الأسطر التي تحوي قيم None

من الرسم البياني الذي يمثل توزع قيم رواتب الموظفين نلاحظ أن معظم الرواتب تتجمع قيمتها في المجال أقل من 100000

ملاحظة : لا يوجد في البيانات عمود department لإظهار توزيع الموظفين على الأقسام

تم تجميع البيانات وفق العمود Year وإيجاد القيم الإحصائية لكل مجموعة والتي تظهر تقارب في متوسط الرواتب بين المجموعات

من الرسم البياني للعلاقة بين ال salary و Benefits نجد أنها علاقة تزايدية