**OOO “Нетология”**

**Курс “Аналитик данных”**

**Выпускная квалификационная работа на тему: “*Анализ заработных плат сотрудников в Сан- Франциско (поиск инсайтов, составление рекомендаций стейкхолдерам)*”**

Автор: Полякова Елена Владимировна

Курс DA-53

2023

**Оглавление**

1. [Введение](#_9wptcbsc4kvd)
2. [Исходные данные](#_rfv1lkwf3fkq)
3. [Анализ данных](#_hyst5ih5h3s0)
4. [Методика решения](#_y3ouir4lp7m3)
5. [Результаты](#_2lj0c7aj0u79)

# Введение

В данной работе проводится анализ данных по заработной плате государственных служащих Сан-Франциско с 2011 по 2018 год.

Данное исследование направлено для анализа заработных плат по годам, по специальностям, выявить самые высоко и низко оплачиваемые позиции.

Оно может быть интересно избирателям, налогоплательщикам - на что тратятся налоговые деньги, лицам в правительстве, в принятии обоснованных решений , HR менеджерам при поисках кандидатов, людям, которые хотели бы найти работу на государственной службе в Сан-Франциско.

Датасет может быть использован для анализа:

1. заработной платы (средняя и медиана) по годам
2. анализ доходов по специальности, по специальности и по году, рост доходов
3. отрицательные виды выплат. Почему они есть и какие специалисты их получают
4. кто получает самые большие и самые низкие зарплаты
5. кто больше получает за переработку, больше перерабатывает
6. в какой специальности большие поощрения (‘Benefits’)
7. действительно ли есть ли рост зарплаты по годам по различным выплатам
8. корреляция между разными видами выплат

Аудитория: общество, общественные организации, правительство, избиратели.

Стейкхолдеры: правительство, избиратели, партии, налогоплательщики

Метрика: заработная плата сотрудников

Гипотезы для проверки:

1. 10 процентов самых высокооплачиваемых сотрудников получают суммарно больше, чем остальные 90.
2. Происходит ли рост заработных плат по годам.

# Исходные данные

Датасет: Salaries for San Francisco Employee

Датасет взят из открытого источника:

[Salaries for San Francisco Employee](https://www.kaggle.com/datasets/mojtaba142/20112018-salaries-for-san-francisco)

Этот набор данных содержит более 300 тысяч записей о заработной плате государственных служащих в Сан-Франциско с 2011 по 2018 год.

Датасет содержит 9 полей.

| Поле | Описание | Число ненулевых записей | Тип поля |
| --- | --- | --- | --- |
| EmployeeName | ФИО Работника | 312882 | object |
| JobTitle | Название должности | 312882 | object |
| BasePay | Основная оплата (оклад) | 312882 | object |
| OvertimePay | Оплата за работу в сверхурочное время | 312882 | object |
| OtherPay | Другие виды оплат | 312882 | object |
| Benefits | Премии | 312882 | object |
| TotalPay | Общая заработная плата | 312882 | object |
| TotalPayBenefits | Общая заработная плата с премиями | 312882 | float64 |
| Year | Год выплаты заработной платы | 312882 | int64 |

Датасет содержит 312882 записей

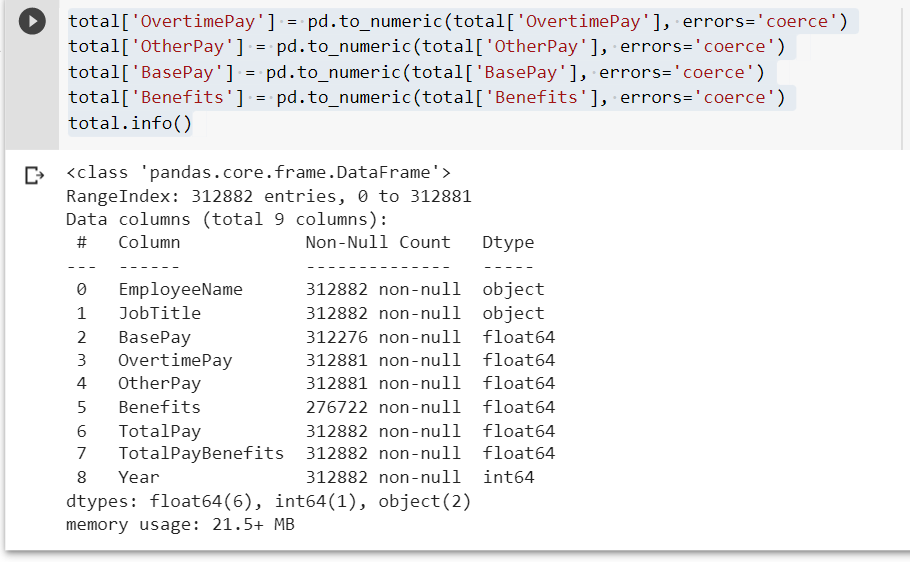
Среда для выполнения работы использована среда Google collab Python

# 

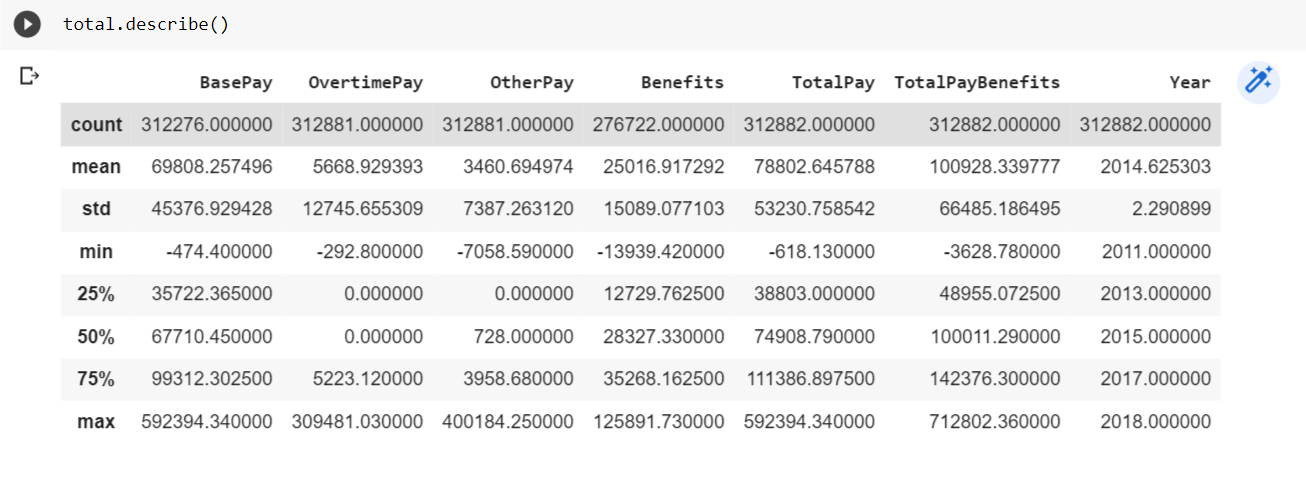
# 

# 3. Анализ данных

Датасет хорошо заполнен, но есть данные, которые не являются числовыми. Их можно либо сделать как неопределенные и проводить анализ с неопределенными данными либо заполнить нулевыми значениями. Поле “Benefits” содержить достаточно много значений “No provided”- данные не предоставлены. Делаем замену как данные не определены, чтобы не получить искажений.

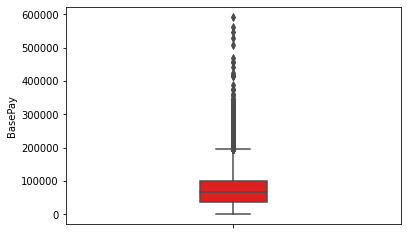


После очистки датасета получены все статистические характеристики датасета. (среднее значение по полям, медианы, максимум и минимум, квантили).

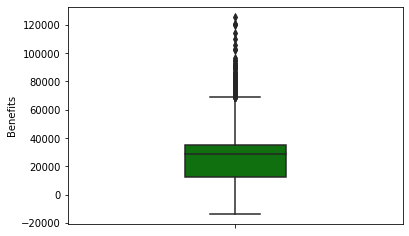


По основным характеристиками мы можем сделать выводы по средним зарплатам, размаху видов оплат и увидеть отклонения.

Посмотрим на графиках, как выглядят данные:

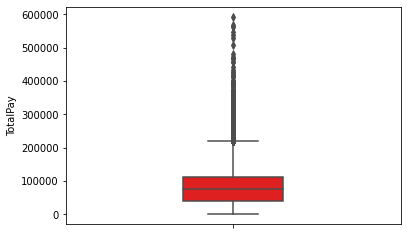


Основная оплата - от 35 до 100 тысяч.



Benefits

и общая оплата



Оценка выбросов. Выбросы есть- в основном это большие значения заработных плат ведущих топ менеджеров, то есть это не ошибочные значения, а просто большие заработные платы. Стоит ли их отбрасывать для анализа данных? Я думаю, нет. Но можно оценить разницу в средних значениях и медианах в полном датасете и в учетом отсечения предполагаемых выбросов.

Кроме этого можно предположить, что выбросами являются и отрицательные значения выплат. Но я думаю, что это не так, и не решусь отсекать эти позиции. Думаю, что их можно исследовать, посмотреть, какие специалисты получают отрицательные типы каких выплат. Возможно, что это штрафы, недоработки, корректировка более ранних начислений, которые не должны быть, ссуды.

Заработные платы могут быть проанализированы по годам, исследован рост заработных плат по годам. Исследовать по годам минимальные и максимальные вознаграждения.

Датасет позволяет найти самые высокооплачиваемые должности и самые низкооплачиваемые.

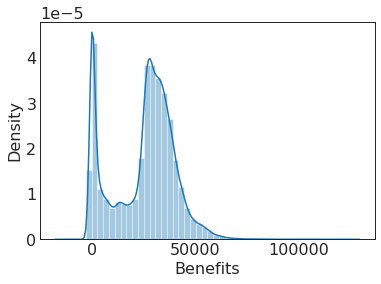
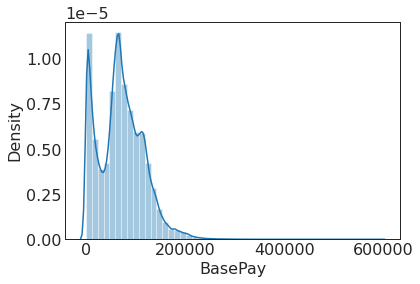
Можно посмотреть корреляцию между различными видами оплат

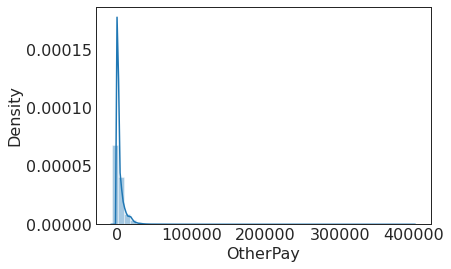
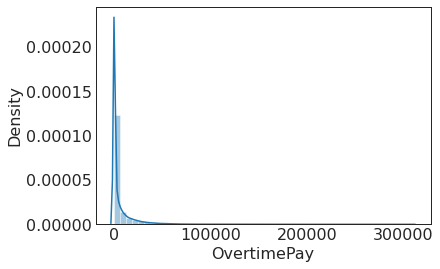
Выбрать наиболее востребованные специальности.

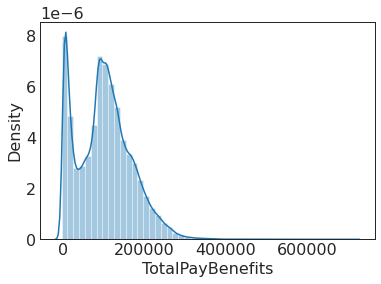
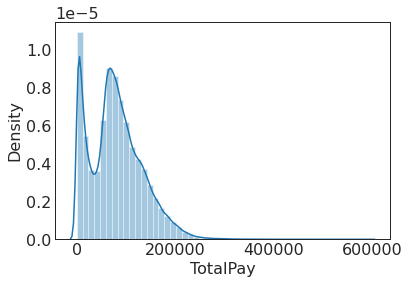
Кроме этого мы можем найти заработную плату любого человека по имени.

# 4. Методика решения

Визуализируем данные по каждому полю и посмотрим распределения.







Мы видим, что распределения не нормальные (хотя OvertimePay и OtherPay немного похожи на нормальные). Распределения по другим видам оплат похожи, имеют большой скачок около 0, потом спад и примерно нормальной распределение, с длинным тонким хвостом в больших значениях.

Сделаем группировку по профессиям и посмотрим средние зарплаты по профессиям. Посмотрим ТОП самых высокооплачиваемых и низкооплачиваемых профессий

| Должность | Заработная плата |
| --- | --- |
| Chief Investment Officer | 611728.63 |
| Physician Administrator, DPH | 454471.81 |
| Managing Director | 438767.56 |
| Chief, Fire Department | 417767.86 |
| GENERAL MANAGER -METROPOLITAN TRANSIT AUTHORITY | 399211.28 |
| Gen Mgr, Public Trnsp Dept | 389126.94 |
| Assistant Chief of Police | 387003.87 |
| Administrator, DPH | 374311.56 |
| Dept Head V | 365736.80 |
| Chief of Police | 360199.83 |

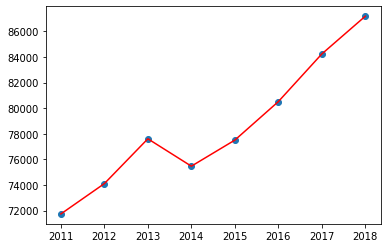
Выберем специальности, которые имеют отрицательные начисления по “BasePay”

Это Junior Clerk, Clerk, PS Aide Health Services, Custodian.

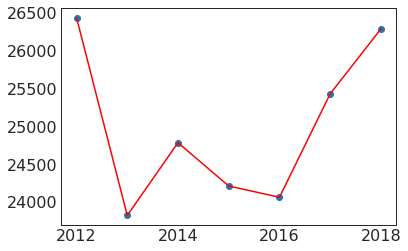
Аналогично можно посмотреть те специальности, где получают отрицательные “Benefits” и итоговую заработную плату.

Проанализируем рост заработных плат по годам- сделаем группировку по году и возьмем среднюю заработную плату.

Получим график по годам. Из него видно, что в 2014 году идет спад средней заработной платы (скорее всего из -за кризиса) и потом рост, в 2015 году зарплата возвращается к уровню 2013 года, далее идет рост

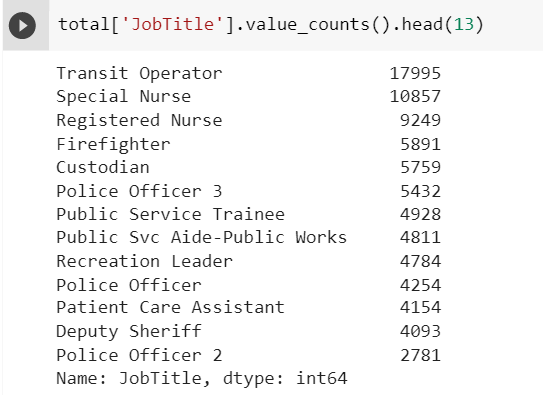


Если рассматривать такую составляющую заработной платы как Benefits, то мы получаем следующий график:

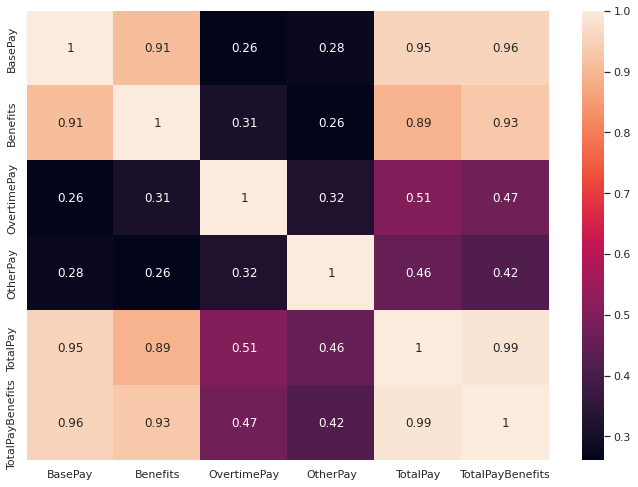


Можно увидеть , что выплаты были максимальными в 2012 году, в 2013 - минимальные.

Получим статистику наиболее востребованных специальностей: наибольшее число работников необходимо



Карта корреляции между разными видами оплат будет выглядеть так:



# 5. Результаты

Результатом этой работы можно считать аналитический отчет по датасету.

Данные датасета были очищены. Данные по заработной плате были преобразованы из Object во float64. Те данные, которые не удалось преобразовать или они содержали информацию 'Not provided' были отброшены.

Получены стандартные статистические характеристики данных по заработной плате, медиана, изучено распределение данных, построены графики. Данные не имеют нормального распределения. Кроме того, имеют какие-то средние значения и небольшой процент очень высоких выплат, что видно на графиках. Можно ли считать, что это выбросы? Нет, нельзя, это высокие зарплаты руководящих должностей.

Кроме того, можно так же посмотреть заработные платы по годам и профессиям и сравнить как со временем меняются заработные платы по профессии.

Составлен ТОП самых востребованных специальностей.

Получены специальности, у которых есть отрицательные выплаты. Являются ли отрицательные выплаты выбросами? Вероятно нет,возможно, это штрафы и просто удалить их или заменить на нулевое значение нельзя.

Составлена карта корреляции между величинами зарплаты. Из нее можно сделать вывод, что BasePay, TotalPay, TotalPayBenefits коррелируют друг с другом, что очевидно так как поля TotalPay, TotalPayBenefits состоят из остальных полей, основную часть которых составляет BasePay, Премия тоже сильно коррелирует со значением BasePay, а поля OtherPay, OvertimePay не коррелируют с другими и друг другом.

Проверка гипотез.

Была проверена гипотеза, что суммарно 10% самых высокооплачиваемых специалистов получают суммарную заработную плату выше всех остальных.

Гипотеза оказалась неверна. Опытным путем получили, что 28 % самых высокооплачиваемых специалистов в сумме получают заработную плату суммарно выше всех остальных 72 %.

Получены списки самых высокооплачиваемых и самых низкооплачиваемых должностей. ТОП 10 самых высоких заработных плат на 415% выше средней по датасету.

Гипотеза о росте средних значений заработной платы по годам не подтвердилась. После исследования средних значений заработных плат по годам, мы увидели, что в 2014 году был спад (вероятно, повлиял кризис), потом пошел рост, в 2015 году заработные платы стали примерно как в 2013 и далее до 2018 шел рост.