SAN FRANCISCO

SALARY ANALYSIS

Presented by: Arpita Chakroborty

Project Objective

- Salary Data Transformation & Analysis Using Pandas and MySQL

- The objective of this project is to leverage both Python's Pandas library for data transformation and MySQL for in-depth data analysis of employee salary records. This dual approach allows for the efficient handling, cleaning, and transforming of raw data with Pandas, while using MySQL to perform more complex analytical queries and generate actionable insights.

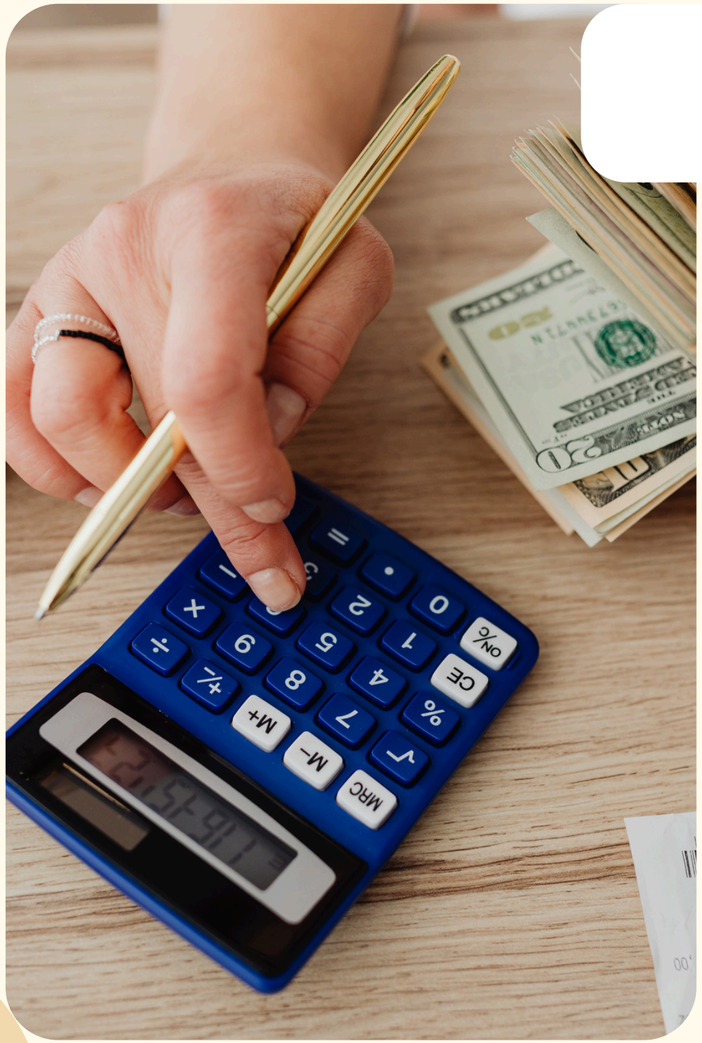
Objectives

Data Cleaning & Transformation (Pandas):

- Clean and preprocess the salary data by handling missing values, formatting issues, and outliers.
- Transform key columns (e.g., base pay, overtime pay, and total pay) to ensure consistency and usability for further analysis.
- Create new calculated fields, such as total compensation, overtime as a percentage of base salary, etc.
- Reshape the data to analyze salary trends over different time periods.

Data Storage & Querying (MySQL):

- Import the transformed data into a MySQL database.
- Use SQL queries to analyze various aspects of the data, such as identifying salary trends, examining salary disparities across job titles, departments, and locations, and determining factors influencing overtime and bonus payments.
- Generate insights regarding salary distribution, high/low earning departments, and yearly salary trends.



Key Insights:

- Identify the highest-paid job roles and departments.
- Analyze the growth of salaries over the years and detect any anomalies or trends.
- Investigate the influence of overtime and benefits on the total compensation package.
- Derive insights into salary disparities and their potential causes (e.g., job title, department, location).

Import libraries and load the data

```
import pandas as pd
```

```
data=pd.read_csv('D:\\DATA ANALYST PROJECT\\resume-project-data-analytics\\Salaries.csv',low_memory=False)
```

```
data.head()
```

	Id	EmployeeName	JobTitle	BasePay	OvertimePay	OtherPay	Benefits	TotalPay	TotalPayBenefits	Year	Notes	Agency	Status
0	1	NATHANIEL FORD	GENERAL MANAGER-METROPOLITAN TRANSIT AUTHORITY	167411.18	0.0	400184.25	NaN	567595.43	567595.43	2011	NaN	San Francisco	NaN
1	2	GARY JIMENEZ	CAPTAIN III (POLICE DEPARTMENT)	155966.02	245131.88	137811.38	NaN	538909.28	538909.28	2011	NaN	San Francisco	NaN
2	3	ALBERT PARDINI	CAPTAIN III (POLICE DEPARTMENT)	212739.13	106088.18	16452.6	NaN	335279.91	335279.91	2011	NaN	San Francisco	NaN
3	4	CHRISTOPHER CHONG	WIRE ROPE CABLE MAINTENANCE MECHANIC	77916.0	56120.71	198306.9	NaN	332343.61	332343.61	2011	NaN	San Francisco	NaN
4	5	PATRICK GARDNER	DEPUTY CHIEF OF DEPARTMENT,(FIRE DEPARTMENT)	134401.6	9737.0	182234.59	NaN	326373.19	326373.19	2011	NaN	San Francisco	NaN

Check last 10 rows of the data set

[7]:
```
# 2.Check last 10 rows of the data set
data.tail(10)
```

[7]:

	Id	EmployeeName	JobTitle	BasePay	OvertimePay	OtherPay	Benefits	TotalPay	TotalPayBenefits	Year	Notes	Agency	Status
148644	148645	Randy D Winn	Stationary Eng, Sewage Plant	0.00	0.00	0.00	0.00	0.00	0.00	2014	NaN	San Francisco	PT
148645	148646	Carolyn A Wilson	Human Services Technician	0.00	0.00	0.00	0.00	0.00	0.00	2014	NaN	San Francisco	PT
148646	148647	Not provided	Not provided	Not Provided	Not Provided	Not Provided	Not Provided	0.00	0.00	2014	NaN	San Francisco	NaN
148647	148648	Joann Anderson	Communications Dispatcher 2	0.00	0.00	0.00	0.00	0.00	0.00	2014	NaN	San Francisco	PT
148648	148649	Leon Walker	Custodian	0.00	0.00	0.00	0.00	0.00	0.00	2014	NaN	San Francisco	PT
148649	148650	Roy I Tillery	Custodian	0.00	0.00	0.00	0.00	0.00	0.00	2014	NaN	San Francisco	PT
148650	148651	Not provided	Not provided	Not Provided	Not Provided	Not Provided	Not Provided	0.00	0.00	2014	NaN	San Francisco	NaN
148651	148652	Not provided	Not provided	Not Provided	Not Provided	Not Provided	Not Provided	0.00	0.00	2014	NaN	San Francisco	NaN
148652	148653	Not provided	Not provided	Not Provided	Not Provided	Not Provided	Not Provided	0.00	0.00	2014	NaN	San Francisco	NaN
148653	148654	Joe Lopez	Counselor, Log Cabin Ranch	0.00	0.00	-618.13	0.00	-618.13	-618.13	2014	NaN	San Francisco	PT

Transform the data

Identifying Non-Numeric Values

```
non_numeric = data[pd.to_numeric(data['Benefits'], errors='coerce').isna()]
print(non_numeric['Benefits'])
```

```
Series([], Name: Benefits, dtype: float64)
```

Fill non-numeric values with a default value (like 0 or mean)

```
# Option 2: Fill non-numeric values with a default value (like 0 or mean)
data['Benefits'] = pd.to_numeric(data['Benefits'], errors='coerce').fillna(0)
```

Change the data type

```
data['Benefits'] = data['Benefits'].astype(float)
print(data.dtypes)
```

```
EmployeeName        object
JobTitle            object
BasePay            float64
OvertimePay        float64
OtherPay           float64
Benefits           float64
TotalPay           float64
TotalPayBenefits   float64
Year                 int64
dtype: object
```

check null values in the data set

```
]: # 5.check null values in the data set
   data.isnull().sum()
```

```
]: Id                       0
   EmployeeName             0
   JobTitle                 0
   BasePay                  0
   OvertimePay              0
   OtherPay                 0
   Benefits                 0
   TotalPay                 0
   TotalPayBenefits         0
   Year                     0
   Notes               148654
   Agency                   0
   Status              110535
   dtype: int64
```

.Drop id ,notes,agency and status columns

```
# 6.Drop id ,notes,agency and status columns
data.columns
```

```
Index(['Id', 'EmployeeName', 'JobTitle', 'BasePay', 'OvertimePay', 'OtherPay',
       'Benefits', 'TotalPay', 'TotalPayBenefits', 'Year', 'Notes', 'Agency',
       'Status'],
      dtype='object')
```

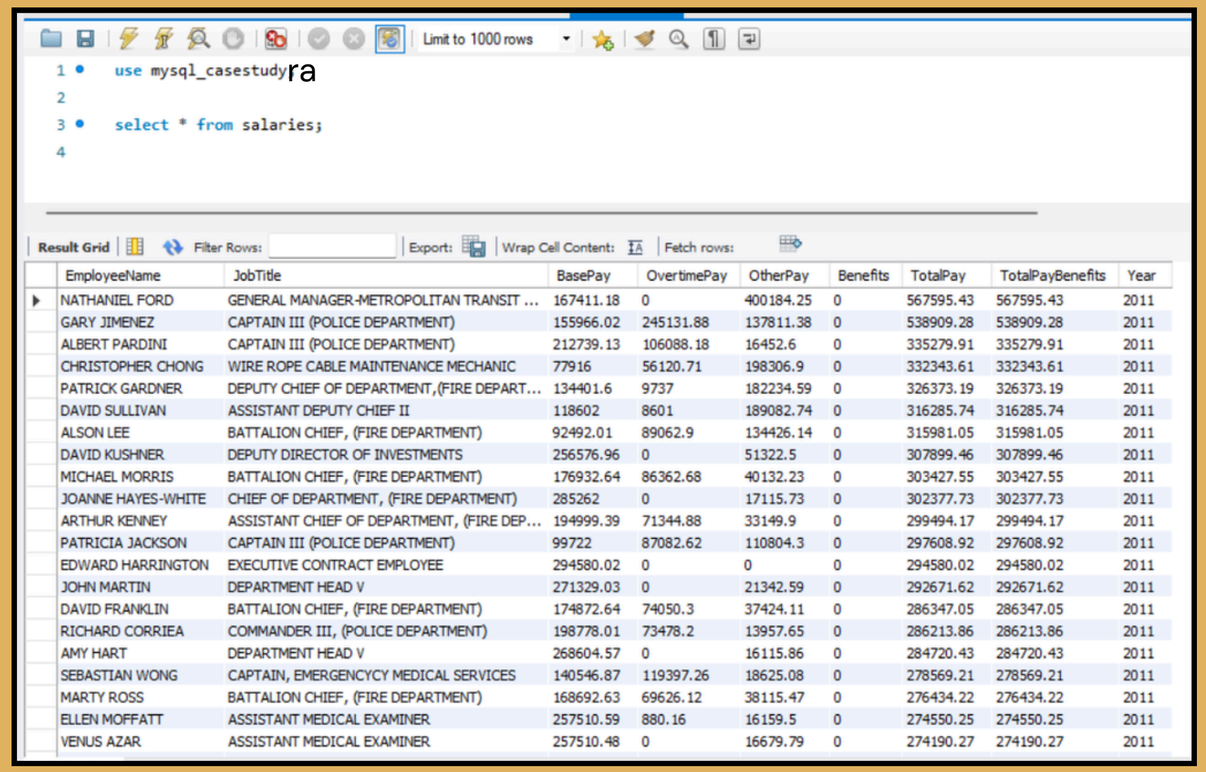
```
data=data.drop(['Id','Notes','Agency','Status'],axis=1)
```

After Transformation Import Data In Mysql

```python
from sqlalchemy import create_engine

# Define connection string
conn_string = 'mysql+pymysql://root:myNewPass123!@localhost/mysql_casestudy'

# Create engine
db = create_engine(conn_string)

# Use engine instead of connection object in to_sql
data.to_sql('salaries', con=db, if_exists='replace', index=False)
```

```
148654
```

Load Data



```
use mysql_casestudyra

select * from salaries;
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content: | Fetch rows:

EmployeeName	JobTitle	BasePay	OvertimePay	OtherPay	Benefits	TotalPay	TotalPayBenefits	Year
NATHANIEL FORD	GENERAL MANAGER-METROPOLITAN TRANSIT ...	167411.18	0	400184.25	0	567595.43	567595.43	2011
GARY JIMENEZ	CAPTAIN III (POLICE DEPARTMENT)	155966.02	245131.88	137811.38	0	538909.28	538909.28	2011
ALBERT PARDINI	CAPTAIN III (POLICE DEPARTMENT)	212739.13	106088.18	16452.6	0	335279.91	335279.91	2011
CHRISTOPHER CHONG	WIRE ROPE CABLE MAINTENANCE MECHANIC	77916	56120.71	198306.9	0	332343.61	332343.61	2011
PATRICK GARDNER	DEPUTY CHIEF OF DEPARTMENT,(FIRE DEPART...	134401.6	9737	182234.59	0	326373.19	326373.19	2011
DAVID SULLIVAN	ASSISTANT DEPUTY CHIEF II	118602	8601	189082.74	0	316285.74	316285.74	2011
ALSON LEE	BATTALION CHIEF, (FIRE DEPARTMENT)	92492.01	89062.9	134426.14	0	315981.05	315981.05	2011
DAVID KUSHNER	DEPUTY DIRECTOR OF INVESTMENTS	256576.96	0	51322.5	0	307899.46	307899.46	2011
MICHAEL MORRIS	BATTALION CHIEF, (FIRE DEPARTMENT)	176932.64	86362.68	40132.23	0	303427.55	303427.55	2011
JOANNE HAYES-WHITE	CHIEF OF DEPARTMENT, (FIRE DEPARTMENT)	285262	0	17115.73	0	302377.73	302377.73	2011
ARTHUR KENNEY	ASSISTANT CHIEF OF DEPARTMENT, (FIRE DEP...	194999.39	71344.88	33149.9	0	299494.17	299494.17	2011
PATRICIA JACKSON	CAPTAIN III (POLICE DEPARTMENT)	99722	87082.62	110804.3	0	297608.92	297608.92	2011
EDWARD HARRINGTON	EXECUTIVE CONTRACT EMPLOYEE	294580.02	0	0	0	294580.02	294580.02	2011
JOHN MARTIN	DEPARTMENT HEAD V	271329.03	0	21342.59	0	292671.62	292671.62	2011
DAVID FRANKLIN	BATTALION CHIEF, (FIRE DEPARTMENT)	174872.64	74050.3	37424.11	0	286347.05	286347.05	2011
RICHARD CORRIEA	COMMANDER III, (POLICE DEPARTMENT)	198778.01	73478.2	13957.65	0	286213.86	286213.86	2011
AMY HART	DEPARTMENT HEAD V	268604.57	0	16115.86	0	284720.43	284720.43	2011
SEBASTIAN WONG	CAPTAIN, EMERGENCYCY MEDICAL SERVICES	140546.87	119397.26	18625.08	0	278569.21	278569.21	2011
MARTY ROSS	BATTALION CHIEF, (FIRE DEPARTMENT)	168692.63	69626.12	38115.47	0	276434.22	276434.22	2011
ELLEN MOFFATT	ASSISTANT MEDICAL EXAMINER	257510.59	880.16	16159.5	0	274550.25	274550.25	2011
VENUS AZAR	ASSISTANT MEDICAL EXAMINER	257510.48	0	16679.79	0	274190.27	274190.27	2011

salary analysis:

General Salary Distribution:

1.What is the distribution of base pay, overtime pay, and other pay across different job titles?

```
select jobtitle,
avg(basepay) avgbasepay,max(basepay) as maximum_basepay,min(basepay) as minimum_basepay,
avg(OvertimePay) avg_OvertimePay ,max(OvertimePay) as maximum_OvertimePay,min(OvertimePay) as minimum_OvertimePay,
avg(OtherPay) avg_OtherPay,max(OtherPay) as maximum_OtherPay,min(OtherPay) as minimum_OtherPay
from salaries
group by jobtitle
order by jobtitle;
```

2. What is the average salary (BasePay, TotalPay, TotalPayBenefits) for different departments or agencies?

```
5 •   select jobtitle as department,
6     round((basepay+overtimepay+otherpay)/3,2)as average_salary
7     from salaries
```

esult Grid | Filter Rows: | Export: | Wrap Cell Content: | Fetc

department	average_salary
GENERAL MANAGER-METROPOLITAN TRANSIT ...	189198.48
CAPTAIN III (POLICE DEPARTMENT)	179636.43
CAPTAIN III (POLICE DEPARTMENT)	111759.97
WIRE ROPE CABLE MAINTENANCE MECHANIC	110781.2
DEPUTY CHIEF OF DEPARTMENT,(FIRE DEPART...	108791.06
ASSISTANT DEPUTY CHIEF II	105428.58

salary analysis:

Salary Trends Over the Years:

How do base pay and total pay vary across different years?

```
7 •  select year ,sum(basepay)as total_basepay,sum(totalpay) as total_pay ,
8    abs(round(sum(basepay) - sum(totalpay),2)) as differce
9    from salaries
10   group by year
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

year	total_basepay	total_pay	differce
2011	2299566191.69005	2594195051.880046	294628860.19
2012	2405834934.5199976	2724848200.4400406	319013265.92
2013	2576380748.040021	2918655930.8000846	342275182.76
2014	2537369199.3399897	2876910951.2599883	339541751.92

Which job titles have the highest and lowest total pay?

```
10 •  (select jobtitle, max(totalpay) as totalpay
11    from salaries
12    group by jobtitle
13    order by   max(totalpay) desc
14    limit 1)
15    union all
16    (select jobtitle, min(totalpay) as totalpay
17    from salaries
18    group by jobtitle
19    order by   min(totalpay)
20    limit 1)
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

jobtitle	totalpay
GENERAL MANAGER-METROPOLITAN TRANSIT ...	567595.43
COUNSELOR, LOG CABIN RANCH	-618.13

salary analysis:

Who are the top 10 highest-paid employees, and what are their job titles?

```
14 •    select employeename,jobtitle, max(totalpay) as highestpaidemployee
15      from salaries
16      group by  employeename,jobtitle
17      order by highestpaidemployee desc
18      limit 10
19
20
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content: | Fetch rows:

employeename	jobtitle	highestpaidemployee
NATHANIEL FORD	GENERAL MANAGER-METROPOLITAN TRANSIT ...	567595.43
GARY JIMENEZ	CAPTAIN III (POLICE DEPARTMENT)	538909.28
David Shinn	Deputy Chief 3	471952.64
Amy P Hart	Asst Med Examiner	390111.98
Gary Altenberg	Lieutenant, Fire Suppression	362844.66
John Goldberg	Captain 3	350403.41
Samson Lai	Battalion Chief, Fire Suppress	347102.32
Ellen G Moffatt	Asst Med Examiner	344187.46
William J Coaker Jr.	Chief Investment Officer	339653.7
Gregory P Suhr	Chief of Police	339282.07

How does the presence of benefits affect total pay for employees?

```
7 •   SELECT
8       (AVG(totalpaybenefits) - AVG(totalpay)) / AVG(totalpay) * 100 AS percentage_increase_due_to_benefits
9     FROM
0       salaries;
```

ult Grid | Filter Rows: | Export: | Wrap Cell Content:

percentage_increase_due_to_benefits
25.310495594681733

salary analysis:

.Find average basepay of all employee per year

```
20    select  year,round(avg(basepay) ,2)as avg_basepay
21    from salaries
22    group by year
23
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

year	avg_basepay
2011	63595.96
2012	65436.41
2013	68509.83
2014	66557.44

show all employee with a totalpay between 50000 and 75000

```
24 •  select * from salaries
25    where totalpay between 50000 and 75000;
26
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content: | Fetch rows:

EmployeeName	JobTitle	BasePay	OvertimePay	OtherPay	Benefits	TotalPay	TotalPayBenefits	Year
MAMIKO NAKAMURA	LIBRARIAN I	72464.94	0	2534.68	0	74999.62	74999.62	2011
HEDLEY PRINCE	WHARFINGER II	72681.97	748.05	1568.71	0	74998.73	74998.73	2011
LINDA BARNARD	RECREATION SUPERVISOR	74268.25	0	730.05	0	74998.3	74998.3	2011
ANGELA WHITTAKER	ADMINISTRATIVE ANALYST	74997.84	0	0	0	74997.84	74997.84	2011
JANE CHU	ADMINISTRATIVE ANALYST	74997.81	0	0	0	74997.81	74997.81	2011
NICHOLAS LAVROV	WATER SERVICE INSPECTOR	74321.9	675.6	0	0	74997.5	74997.5	2011
EUNICE CHIBUNDU	MENTAL HEALTH REHABILITATION WORKER	57864.92	16380.91	750	0	74995.83	74995.83	2011
THERESA LOONEY	TRUCK DRIVER	74070.11	0	920.25	0	74990.36	74990.36	2011
STEVEN LARA	MANAGER III	68326.5	0	6663.27	0	74989.77	74989.77	2011
JANE CHEN	DIETITIAN	72844.42	0	2144	0	74988.42	74988.42	2011
MICHELLE JOHNSON	SPECIAL NURSE	67605.7	114.93	7267.24	0	74987.87	74987.87	2011
ARMANDO LOPEZ	TRANSIT OPERATOR	62249.17	11079.07	1649.94	0	74978.18	74978.18	2011
GARY GEE	ASSISTANT CONSTRUCTION INSPECTOR	74275.03	700.53	0	0	74975.56	74975.56	2011
ILANA BERNSTEIN	PHYSICAL THERAPIST	74971.32	0	0	0	74971.32	74971.32	2011

salary analysis:

display all the employee names from fire department

```
27 •  select employeename,jobtitle
28    from salaries
29    where jobtitle like '%fire%'
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content: | Fetch

employeename	jobtitle
ROBERT SERRANO	BATTALION CHIEF, (FIRE DEPARTMENT)
JAMES VANNUCCHI	BATTALION CHIEF, (FIRE DEPARTMENT)
PHILIP STEVENS	CAPTAIN, FIRE SUPPRESSION
TIMOTHY SULLIVAN	FIREFIGHTER
CANTREZ TRIPLETT	LIEUTENANT, FIRE DEPARTMENT
NOEL MORONEY	FIREFIGHTER
GERALD MANSUR JR	FIRE FIGHTER PARAMEDIC

maximum salary of police department at differnt categories

```
50 •  select jobtitle,
51    max(totalpay) as totalpay
52    from salaries
53    group by jobtitle
54    having jobtitle like '%police department%'
55    order by totalpay  desc
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

jobtitle	totalpay
CAPTAIN III (POLICE DEPARTMENT)	538909.28
COMMANDER III, (POLICE DEPARTMENT)	286213.86
DEPUTY CHIEF III (POLICE DEPARTMENT)	264074.6
INSPECTOR III, (POLICE DEPARTMENT)	258588.39
LIEUTENANT III (POLICE DEPARTMENT)	251935.01

salary analysis:

What is the name of highest paid person (including benefits)

```
50 •  select employeename, max(totalpaybenefits) as highestpaid_salary
51    from salaries
52    group by employeename
53    order by highestpaid_salary desc
54    limit 1;
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content: | Fetch rows:

employeename	highestpaid_salary
NATHANIEL FORD	567595.43

what are the top 5 most popular jobs based on their salary?

```
56 •  select  distinct jobtitle from ( select *, row_number() over (partition by jobtitle order by totalpay desc) as rn
57    from salaries) x
58    where x.rn=1
59    limit 5
60
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

jobtitle
Account Clerk
ACCOUNTANT
Accountant I
Accountant II
Accountant III

This project will demonstrate how Pandas and MySQL can work together to perform a complete end-to-end data transformation and analysis, showcasing skills in both data manipulation and database management for real-world salary data.

Thank's For Watching

Arpita Chakroborty