Analyzing Departmental Salary Disparities Project

By: Yong Sook Prasit Attavit

Link to accompanying JupyterNotebook with EDA:
Departmental Salary Disparities Analysis_AccompanyingJupyterNotebook

Link to supporting Tableau Visualization:
Supporting Tableau Visualizations

Problem Statement

Business Case:

- The data analytics manager of a company would like to seek insights into salary disparities present within the company department
    - PWD Department has been flagged as a department that has a high amount of salary spread

Objective:

- Obtain relevant insights with Exploratory Data Analysis (EDA), and create a SQL query that identifies a high amount of variation within the department
- Provide the top 5 department that should be selected for management to review, with regards to having the most variance & discrepancy in salary

Deliverables:

- Provide a list from a SQL database with a way to score variation by Department
- JupyterNotebook with accompanying Python code block for SQL calculation cross-validation & EDA

Dataset Glossary

Field Name	Description
Department	*3 Letter alphabetical code of the department in which the employee belongs to*
Department_Division	*Contains both the departmental alphabetical code and the corresponding division of the employee*
PCN	*Unique identifier or code assigned to each individual employee within an organization's HR system.*
Position_Title	*Title of the position of which the employee holds*
FLSA_Status	*Employee Classified under the Fair Labor Standards Act [FLSA], in which an employee is classified as either a non-exempt employee or an exempt employee*
Initial_Hire_Date	*Initial hire date of employee*
Date_in_Title	*Date which the employee started holding the Position_Title*
Salary	*Salary information of employee*
Hourly_Annual_Salaried Employee ***[Self-Created Column]***	*Self-created categorical column where we eventually define Salary <=10,000 as Hourly Salaried Employee & Salary > 10,000 as Annual Salaried Employee after EDA*

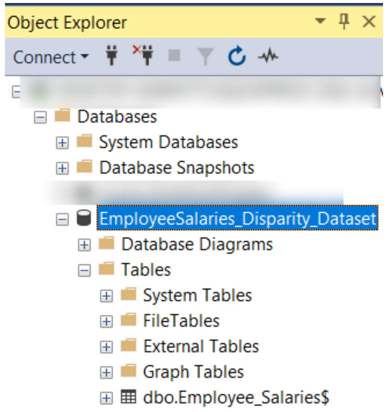
Methodology

Relevant datasets can be obtained via my Github:

Analyzing Departmental Salary Disparities Github Root Directory

- /data contains the original dataset used for analysis which is: Employee_Salaries.csv
- Departmental Salary Disparities Analysis Project_AccompanyingJupyterNotebook.ipynb is the accompanying JupyterNotebook used for EDA and data visualization
    - It focuses on departmental histogram plots, quantile-quantile plots, as well as cross validating SQL calculation
- EmployeeSalaries_Disparity_Dataset.sql contains the SQL codes used in this project

Raw .csv file is ingested into Microsoft SQL Server Management Studio (SSMS) and SQL queries were iteratively built upon to obtain the final output which will identify departmental employee salary variation



Sanity Check with Python and SQL

Overall sanity check on missing/NaN values

- A check on missing data present within the dataset was first done with Python in the accompanying JupyterNotebook:

```
#### Sanity check on missing values within dataset

# Calculate the count of null values for each column
count_nan = df.isnull().sum()

# Calculate the total number of rows in the dataframe
total_rows = len(df)

# Calculate the percentage of null values for each column
percentage_nan = (count_nan / total_rows) * 100

# Combine the count and percentage values side by side
nan_summary = pd.concat([count_nan, percentage_nan], axis=1, keys=['Count of Null', '% of Null'])

print(nan_summary)
```

```
                     Count of Null  % of Null
Department                       0   0.000000
Department_Division              0   0.000000
PCN                              0   0.000000
Position_Title                   0   0.000000
FLSA_Status                     21   0.301984
Initial_Hire_Date                0   0.000000
Date_in_Title                  901  12.956572
Salary                           8   0.115042
```

Here, I observe that null values are present in 'FLSA_Status', 'Date_in_Title' and 'Salary' columns

'Date_in_Title' is not used in the scope of this project & will be thus ignored regarding missing values

Keeping in mind that the focus of this project is to investigate salary disparity, I'll place a greater emphasis on the missing values present in the 'Salary' column

Sanity Check with Python and SQL

'FLSA_Status' sanity check on missing/NaN values

```
## Check for missing values in 'FLSA_Status'
df[df['FLSA_Status'].isnull()]
```

	Department	Department_Division	PCN	Position_Title	FLSA_Status	Initial_Hire_Date	Date_in_Title	Salary
1455	GRD	GRD 010 Voter Registration and Elections	N.030141	Board of Equalization	NaN	1/1/2021	NaN	NaN
1460	GRD	GRD 010 Voter Registration and Elections	N.030141	Board of Equalization	NaN	5/14/2018	NaN	NaN
1468	GRD	GRD 010 Voter Registration and Elections	N.030141	Board of Equalization	NaN	5/15/2018	NaN	NaN
3200	PAR	PAR 070 CRC - Kempsville	N.090153	Lifeguard	NaN	1/9/2020	NaN	12.06
3419	PAR	PAR 073 CRC - Great Neck	N.090161	Aquatics Instructor	NaN	2/5/2020	NaN	15.72
3512	PAR	PAR 075 CRC - Princess Anne	N.090163	Aquatics Instructor	NaN	1/17/2020	NaN	15.72
3540	PAR	PAR 075 CRC - Princess Anne	N.090163	Aquatics Instructor	NaN	3/10/2020	NaN	15.72
3804	PAR	PAR 085 Therapeutic Recreation Programs	N.030440	Activity Center Assistant Leader	NaN	1/27/2020	NaN	12.80
3959	PAR	PAR 089 Out-Of-School Time - School Based	N.030093	Activity Center Assistant Leader	NaN	7/9/2020	NaN	12.80
3961	PAR	PAR 089 Out-Of-School Time - School Based	N.030093	Activity Center Assistant Leader	NaN	7/9/2020	NaN	12.80
3999	PAR	PAR 089 Out-Of-School Time - School Based	N.030094	Activity Center Leader	NaN	7/9/2020	NaN	14.89
6402	SHF	SHF 033 Court Support Services	S.020001	Security Screener (State)	NaN	7/26/2021	7/26/2021	14.13
6408	SHF	SHF 033 Court Support Services	S.020001	Security Screener (State)	NaN	8/19/2021	8/19/2021	14.13
6421	SHF	SHF 033 Court Support Services	S.020001	Security Screener (State)	NaN	9/2/2021	9/2/2021	14.13
6442	SHF	SHF 033 Court Support Services	S.020001	Security Screener (State)	NaN	7/1/2001	6/3/2021	22.78
6452	SHF	SHF 033 Court Support Services	S.020001	Security Screener (State)	NaN	9/30/1976	NaN	22.78
6469	SHF	SHF 033 Court Support Services	S.020001	Security Screener (State)	NaN	2/1/2000	6/3/2021	22.78
6474	SHF	SHF 033 Court Support Services	S.020001	Security Screener (State)	NaN	8/26/2021	8/26/2021	14.13
6477	SHF	SHF 033 Court Support Services	S.020001	Security Screener (State)	NaN	11/16/2012	8/19/2021	14.13
6490	SHF	SHF 033 Court Support Services	S.020001	Security Screener (State)	NaN	2/1/1997	6/17/2021	22.78
6542	SHF	SHF 034 Correctional Operations	S.020066.2	Public Safety Investigator (State)	NaN	7/15/2021	7/15/2021	24.02

The majority of salaries where 'FLSA_Status' is null belong to the Lower Income Bracket range. Since our main analysis is focused on the Upper Income Bracket, these missing values are not likely to significantly impact the final calculations and analysis.

[Explained later in the Powerpoint Slides]

Sanity Check with Python and SQL

'Salary' sanity check on missing/NaN values

```python
## Sanity check for missing salary information
df[df['Salary'].isnull()]
```

	Department	Department_Division	PCN	Position_Title	FLSA_Status	Initial_Hire_Date	Date_in_Title	Salary
1455	GRD	GRD 010 Voter Registration and Elections	N.030141	Board of Equalization	NaN	1/1/2021	NaN	NaN
1458	GRD	GRD 010 Voter Registration and Elections	N.030141	Board of Equalization	Exempt	3/1/2014	3/1/2014	NaN
1460	GRD	GRD 010 Voter Registration and Elections	N.030141	Board of Equalization	NaN	5/14/2018	NaN	NaN
1466	GRD	GRD 010 Voter Registration and Elections	N.030141	Board of Equalization	Exempt	1/1/2007	NaN	NaN
1468	GRD	GRD 010 Voter Registration and Elections	N.030141	Board of Equalization	NaN	5/15/2018	NaN	NaN
6321	REA	REA 011 Board of Equalization	N.030010	Board of Equalization	Exempt	8/1/2015	NaN	NaN
6322	REA	REA 011 Board of Equalization	N.030010	Board of Equalization	Exempt	7/1/2016	NaN	NaN
6323	REA	REA 011 Board of Equalization	N.030010	Board of Equalization	Exempt	7/1/2012	NaN	NaN

	Count of Null	% of Null
Department	0	0.000000
Department_Division	0	0.000000
PCN	0	0.000000
Position_Title	0	0.000000
FLSA_Status	21	0.301984
Initial_Hire_Date	0	0.000000
Date_in_Title	901	12.956572
Salary	8	0.115042

- Missing 'Salary' data accounts for <0.12% of the total column
- These missing 'Salary' values corresponds to 2 unique PCN ID corresponding to 'N.030141' & 'N.030010'
- A possible workaround is to check with data engineering team/ data analytics manager to request for salary data for these employees for more conclusive analysis

We'll proceed with data analysis without any missing salary value imputation

Sanity Check with Python and SQL

Sanity check on salary validity for duplicated PCN IDs

- I wanted to check if salary information for duplicated PCN IDs are present, and if so, are the corresponding salary information keyed in sensibly
  - i.e for the same unique ‘PCN’ ID & ‘Position_Title’, the salary should be listed as similar values without much deviation of one another

```sql
-- Sanity Check for duplicates within PCN. Duplicates are present within dataset but duplicate PCN were
confirmed to have mostly same salary values & the same department for duplicate PCN rows. This means
that we can proceed with building our SQL query without much worry. I.E. In a 'dirtier' dataset, the
same unique PCN ID employee might be incorrectly listed that he/she is present in 2/3 more departments
without changing position title and having a wide spread of salary range.
SELECT PCN, Department, Department_Division, Position_Title, FLSA_Status, Initial_Hire_date,
Date_in_Title, Salary, COUNT(PCN) OVER (PARTITION BY PCN) AS Count_of_PCN_ID
FROM EmployeeSalaries_Disparity_Dataset.dbo.Employee_Salaries$
GROUP BY PCN, Department, Department_Division, Position_Title, FLSA_Status, Initial_Hire_date,
Date_in_Title, Salary
HAVING COUNT(PCN) > 1
ORDER BY PCN
```

	PCN	Department	Department_Division	Position_Title	FLSA_Status	Initial_Hire_date	Date_in_Title	Salary	Count_of_PCN_ID
38	N.030713	HSD	HSD 501 Virginia Beach Juvenile Detention Ce...	Juvenile Detention Counselor	Non Exempt	2006-11-01 00:00:00.000	2006-11-01 00:00:00.000	21.21	3
39	N.030751	PWD	PWD 332 WM Bureau of Waste Collection	Waste Management Operator I	Non Exempt	2021-04-08 00:00:00.000	NULL	13.06	1
40	N.090045	EMS	EMS 060 Lifeguard Services	Beach Lifeguard Supervisor	Non Exempt	2015-05-21 00:00:00.000	NULL	15.98	2
41	N.090045	EMS	EMS 060 Lifeguard Services	Beach Lifeguard Supervisor	Non Exempt	2016-05-26 00:00:00.000	NULL	15.98	2
42	N.090046	EMS	EMS 060 Lifeguard Services	Beach Lifeguard	Non Exempt	2017-05-25 00:00:00.000	NULL	13.33	8
43	N.090046	EMS	EMS 060 Lifeguard Services	Beach Lifeguard	Non Exempt	2018-06-07 00:00:00.000	NULL	13.33	8
44	N.090046	EMS	EMS 060 Lifeguard Services	Beach Lifeguard	Non Exempt	2019-05-23 00:00:00.000	NULL	13.33	8
45	N.090046	EMS	EMS 060 Lifeguard Services	Beach Lifeguard	Non Exempt	2019-06-13 00:00:00.000	NULL	13.33	8
46	N.090046	EMS	EMS 060 Lifeguard Services	Beach Lifeguard	Non Exempt	2020-05-21 00:00:00.000	NULL	13.33	8
47	N.090046	EMS	EMS 060 Lifeguard Services	Beach Lifeguard	Non Exempt	2020-06-04 00:00:00.000	NULL	13.33	8
48	N.090046	EMS	EMS 060 Lifeguard Services	Beach Lifeguard	Non Exempt	2021-05-20 00:00:00.000	NULL	13.33	8
49	N.090046	EMS	EMS 060 Lifeguard Services	Beach Lifeguard	Non Exempt	2021-06-03 00:00:00.000	NULL	13.33	8

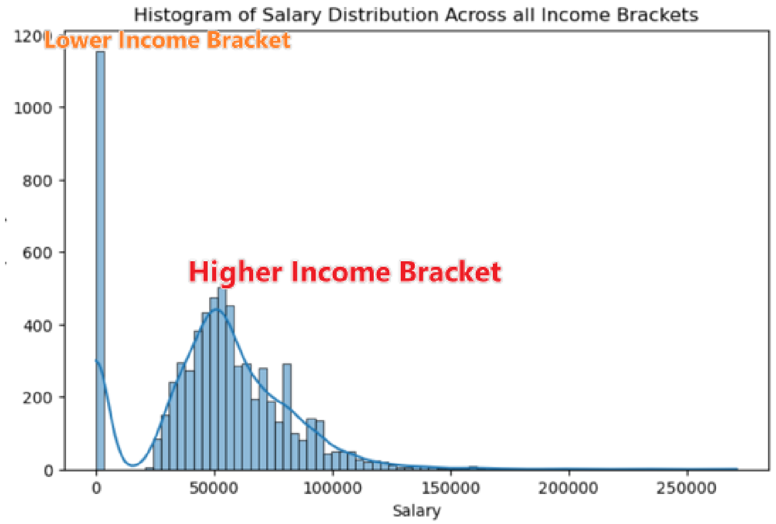
From the observed output, it was observed that rows containing duplicate PCN IDs had mostly repeated salary values or values within the same range

- Hence, the dataset is safe to use for calculation as it will not impact the standard deviation / mean used for the calculation of CV value

Exploratory Data Analysis [EDA] on overall dataset

Initial EDA was performed on the overall dataset using Python in the accompanying JupyterNotebook

- In particular, the EDA focused on the distribution of employee salaries via a Histogram plot

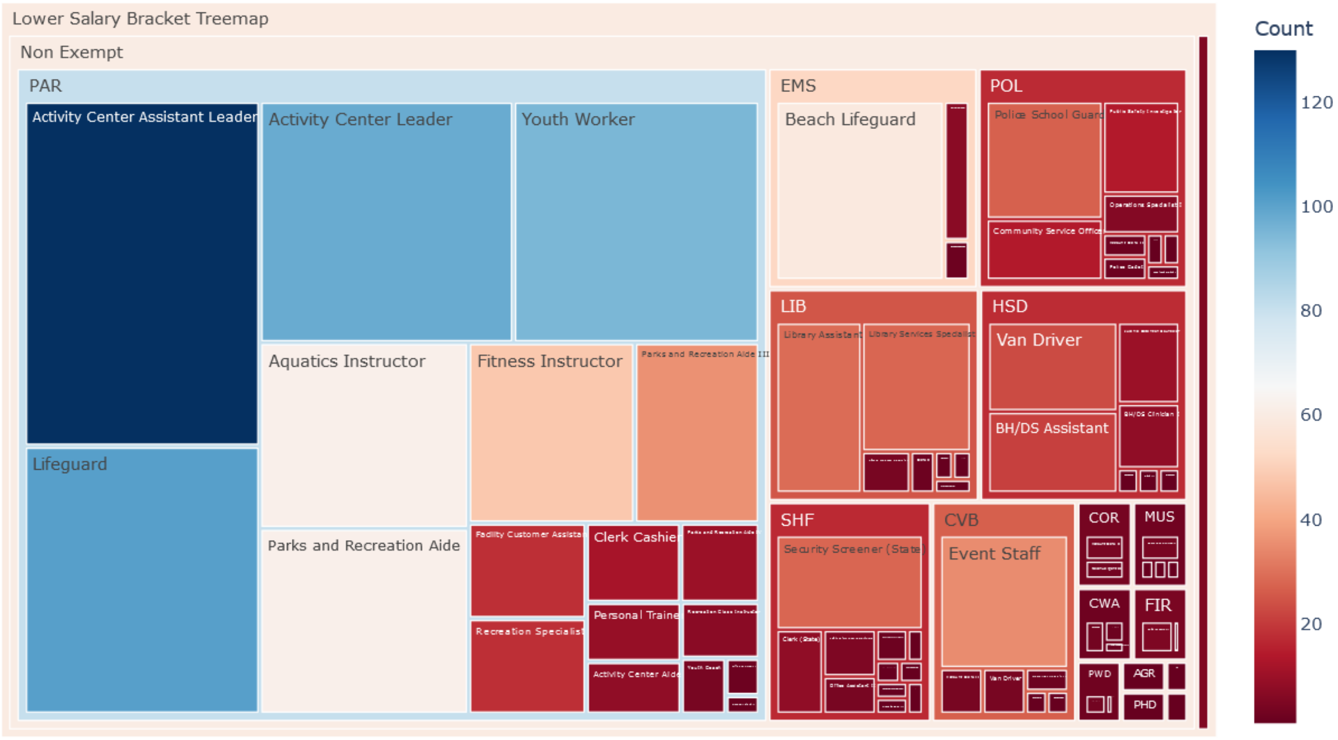


From this plot, I noticed that the highest count of employees within the company fall into the lower income bracket of <= $10,000

EDA on this 'lower income bracket' is performed in the next slide to determine if the 'lower income bracket' should be considered for departmental salary disparity analysis

Exploratory Data Analysis [EDA] on ‘Lower Income Bracket’

Treemap of Lower Income Bracket under FLSA_Status



Upon further investigation, a significant number of employees whose salaries fall in the lower income bracket of <= $10,000 are mainly made up of non-exempt staff (~1124 employees) as compared to exempt staff (~13 employees), and are mostly from the ‘PAR’ Department

According to the Fair Labor Standards Act (FLSA):

- Exempt staff are not eligible for overtime pay and are paid a fixed salary, often performing managerial or professional duties which have typically higher barrier of skill entry
- Non-exempt staff are defined as defined as staff members who are eligible for overtime pay for hours worked beyond 40 per week, and they usually receive hourly wages.
    - As inferred from the Tree-map, it was also observed that the majority of non-exempt staff hold positions that have a lower barrier of skill entry. For example, such as ‘Clerk Cashier’

Following that, I investigated the statistical distribution of the lower income bracket group in JupyterNotebook:

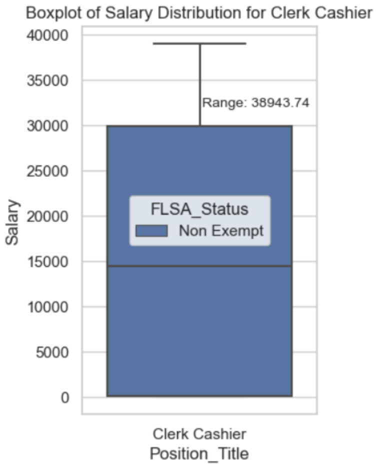
	count	mean	std	min	25%	50%	75%	max
Salary	1155.0	14.25742	4.384736	0.0	11.5	13.33	15.72	51.99

- Here we observe that in the lower income bracket, the mean salary is <$100
- I suspect that most of the employees in the lower income bracket have their salary listed as hourly wage rather than annual wage
- I want to re-examine the characteristics & distribution of employees in this income group using ‘Clerk Cashier’ as an example in the next slide with this assumption in mind

Exploratory Data Analysis [EDA] on ‘Lower Income Bracket’

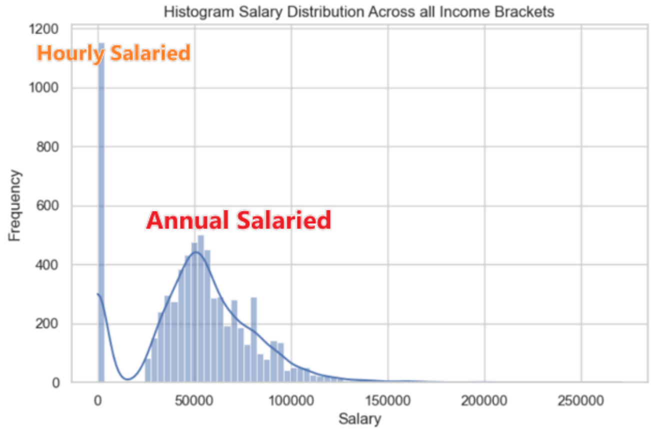
- Taking 'Clerk Cashier' as an example of a non-exempt staff with a lower barrier of skill entry:
- We observe the salary for the position title of 'Clerk Cashier' varies widely according to the box plot plotted and is further supported by the calculated range value and the SQL query output

	Department	Department_Division	PCN	Position_Title	FLSA_Status	Initial_Hire_Date	Date_in_Title	Salary	Hourly_Annual_Salaried_Employee
1	PAR	PAR 073 CRC - Great Neck	B.001888.4	Clerk Cashier	Non Exempt	1997-10-08 00:00:00.000	2001-02-01 00:00:00.000	38957.88	Annual Salaried Employee
2	PAR	PAR 074 CRC - Bayside	B.004803.2	Clerk Cashier	Non Exempt	2013-10-01 00:00:00.000	2016-10-22 00:00:00.000	31532.28	Annual Salaried Employee
3	PAR	PAR 076 CRC - Williams Farm	B.005252.5	Clerk Cashier	Non Exempt	2017-04-27 00:00:00.000	2017-04-27 00:00:00.000	31178.68	Annual Salaried Employee
4	MUS	MUS 028 Aquarium Guest Experiences	B.006709.2	Clerk Cashier	Non Exempt	2017-02-16 00:00:00.000	2017-02-16 00:00:00.000	31178.68	Annual Salaried Employee
5	MUS	MUS 028 Aquarium Guest Experiences	B.007302.1	Clerk Cashier	Non Exempt	2018-01-18 00:00:00.000	2018-01-18 00:00:00.000	31137.08	Annual Salaried Employee
6	PAR	PAR 075 CRC - Princess Anne	B.004545.2	Clerk Cashier	Non Exempt	2018-07-05 00:00:00.000	2018-10-25 00:00:00.000	30450.68	Annual Salaried Employee
7	PAR	PAR 070 CRC - Kempsville	B.007148	Clerk Cashier	Non Exempt	2015-08-05 00:00:00.000	2020-03-12 00:00:00.000	29743.48	Annual Salaried Employee
8	MUS	MUS 028 Aquarium Guest Experiences	B.006710.1	Clerk Cashier	Non Exempt	2018-07-30 00:00:00.000	2019-10-24 00:00:00.000	29743.48	Annual Salaried Employee
9	MUS	MUS 022 Aquarium Exhibits and Technology	B.007297.1	Clerk Cashier	Non Exempt	2021-06-17 00:00:00.000	2021-06-17 00:00:00.000	28849.6	Annual Salaried Employee
10	PAR	PAR 074 CRC - Bayside	B.004801.2	Clerk Cashier	Non Exempt	2017-11-09 00:00:00.000	2021-06-17 00:00:00.000	28849.6	Annual Salaried Employee
11	PAR	PAR 070 CRC - Kempsville	B.007147	Clerk Cashier	Non Exempt	2021-07-08 00:00:00.000	NULL	28849.6	Annual Salaried Employee
12	PAR	PAR 071 CRC - Bow Creek	B.004980.3	Clerk Cashier	Non Exempt	2016-04-13 00:00:00.000	2021-06-17 00:00:00.000	28849.6	Annual Salaried Employee
13	PAR	PAR 071 CRC - Bow Creek	P.050210.1	Clerk Cashier	Non Exempt	2021-07-23 00:00:00.000	NULL	14.14	Hourly Salaried Employee
14	PAR	PAR 072 CRC - Seatack	P.050211.1	Clerk Cashier	Non Exempt	2017-12-13 00:00:00.000	NULL	14.14	Hourly Salaried Employee
15	PAR	PAR 072 CRC - Seatack	P.050128.2	Clerk Cashier	Non Exempt	2021-02-22 00:00:00.000	NULL	14.14	Hourly Salaried Employee
16	PAR	PAR 073 CRC - Great Neck	P.050212.1	Clerk Cashier	Non Exempt	2017-05-30 00:00:00.000	2018-07-19 00:00:00.000	14.14	Hourly Salaried Employee
17	PAR	PAR 073 CRC - Great Neck	P.050129.2	Clerk Cashier	Non Exempt	2015-01-07 00:00:00.000	2016-08-18 00:00:00.000	14.14	Hourly Salaried Employee
18	PAR	PAR 074 CRC - Bayside	P.050125.3	Clerk Cashier	Non Exempt	2021-07-08 00:00:00.000	NULL	14.14	Hourly Salaried Employee
19	PAR	PAR 076 CRC - Williams Farm	P.050185.2	Clerk Cashier	Non Exempt	2021-04-15 00:00:00.000	NULL	14.14	Hourly Salaried Employee
20	PAR	PAR 076 CRC - Williams Farm	P.050215.1	Clerk Cashier	Non Exempt	2021-07-26 00:00:00.000	NULL	14.14	Hourly Salaried Employee
21	PAR	PAR 070 CRC - Kempsville	P.050186.2	Clerk Cashier	Non Exempt	2021-04-29 00:00:00.000	NULL	14.14	Hourly Salaried Employee



Hence, we conclude that clerk cashiers with a lower salary value of ~$14.14 are most likely to be hourly salaried workers and clerk cashiers with a higher salary value of >$10,000 are most likely paid annual wages

It appears that our dataset contains salary information of BOTH hourly and annual waged workers, which are characterized by the 2 distinct peaks in the initial 'Histogram of Salary Distribution Across all Income Brackets' plot.



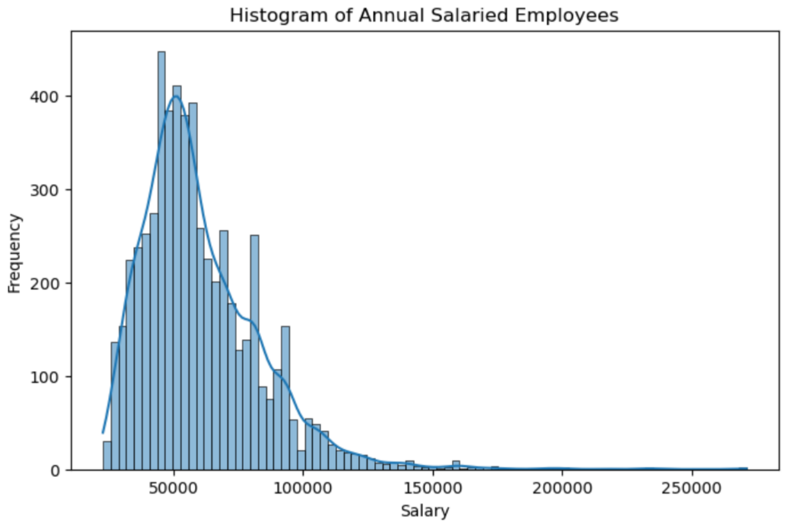
Owing to the fact that it's difficult to accurately predict the actual take-home amount of hourly salaried workers as the # of hours that they've worked is not listed in the dataset, I'd like to limit the departmental salary disparity analysis to annual salaried workers only, which corresponds to Salary >$10,000

From this point on, I’ll refer to 'Lower Income Bracket’ employees as ‘Hourly Salaried Employee' and ‘Higher Income Bracket’ employees as ‘Annual Salaried Employee’.

Hence, in this study, I’ll focus on the investigation of salary disparity amongst departments that contain annual salaried employees only.

Exploratory Data Analysis [EDA] on Annual Salaried Employees

[Previously labelled as Higher Income Bracket]



Statistical Summary of Annual Salaried Employees [>$10,000]			
	Value	**Definition**	**Interpretation**
Sample Size [N]	**5791**	# of sample present in Higher Income Bracket	There are 5791 samples within the Higher Income Bracket
Skewness	**1.6573**	Skewness measures the asymmetry of a distribution.	Our dataset is positively/ right-skewed (skewness > 0). Visually this is indicated with a long tail on the right side of the distribution
Kurtosis	**6.3187**	Kurtosis measures the peakedness or heaviness of the tails of a distribution.	Our dataset follows a leptokurtic distribution, which is characterized with heavy tails and a sharper peak as we have High kurtosis (> 3)
Mean	**60343.70**	Sum of values in the dataset/ # of values in dataset	From the mean and median, we can also deduce that we have a right-skewed distribution as well For a right-skewed distribution, the mean is often greater than the median 
Median	**55224.00**	"Middle" value of the dataset when arranged in ascending or descending order	
Quantile-Quantile [Q-Q] Plot	**N/A**	N/A	Observation: 1. Points do not follow the diagonal [marked in red] • Higher Income Bracket is not normally distributed 2. Point curves upwards • Heavier tails compared to theoretical distribution

Histogram & Quantile-Quantile [Q-Q] plots serves to illustrate the distribution of the data, from the table above, we summarize that for annual salaried workers:

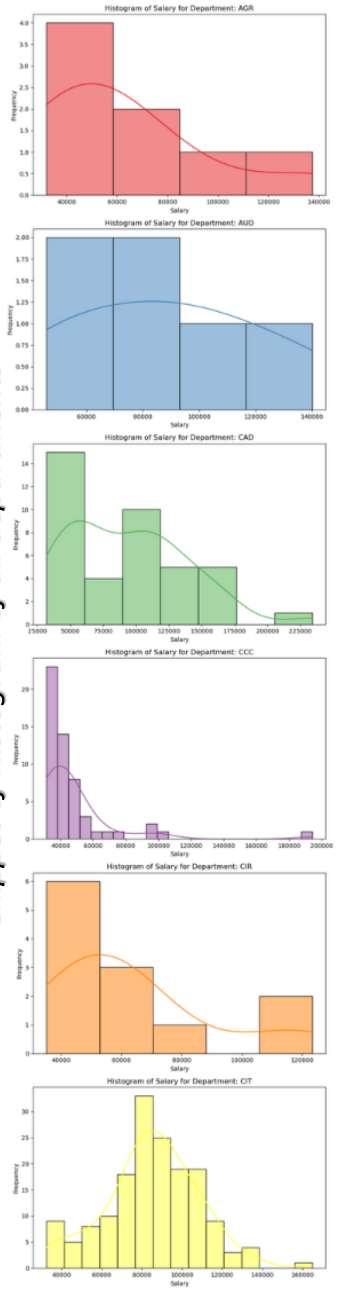
- Our dataset is not normally distributed & is positively skewed/right-skewed

Z-Score values are typically more significant when a dataset is normally distributed

- Taking into consideration that the end goal is to identify departments with salary disparity, I've decided to place more emphasis on the calculated Coefficient of Variation value as opposed to Outlier Counts obtained from Z-Score values on a non-normally distributed dataset

Histogram of all Annual Salaried Employees by Department

Snippet of Histogram of all departments



Python script using for statement to obtain annual salaried workers histogram plot by department

```python
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

df = pd.read_csv('data/Employee_Salaries.csv')
# Filter out rows where 'Salary' is <= 10000
df_salary_filtered = df[df['Salary'] > 10000]
df_salary_filtered['Department'].unique() ## checking departments

# # # Define the list of departments to analyze, in this case I'm inspecting all histograms for annual
salaried workers
department_list_to_analyse = ['AGR', 'AUD', 'CAD', 'CCC', 'CIR', 'CIT', 'CLK', 'CMD', 'COM',
       'COR', 'CUL', 'CVB', 'CWA', 'ECC', 'ECO', 'EMS', 'FIN', 'FIR',
       'GRD', 'HNP', 'HRD', 'HSD', 'JUV', 'LIB', 'MSB', 'MUS', 'OEM',
       'PAR', 'PHD', 'PLN', 'POL', 'PUD', 'PWD', 'REA', 'RMO', 'SHF',
       'STR', 'TRE']

# Filter the DataFrame to include only the departments in department_list_to_analyse
df_salary_filtered_dept_slice =
df_salary_filtered[df_salary_filtered['Department'].isin(department_list_to_analyse)]

# Define a color palette with unique colors for each department
color_palette = sns.color_palette("Set1", n_colors=len(department_list_to_analyse))

# Create a single figure with multiple subplots
fig, axs = plt.subplots(nrows=len(department_list_to_analyse), figsize=(8, 5 *
len(department_list_to_analyse)))

# Loop through each department in department_list_to_analyse and plot a histogram
for i, department in enumerate(department_list_to_analyse):
    sns.histplot(data=df_salary_filtered_dept_slice[df_salary_filtered_dept_slice['Department'] ==
department],
                 x='Salary', kde=True, bins='auto', color=color_palette[i], legend=True, ax=axs[i])
    axs[i].set_title(f"Histogram of Salary for Department: {department}")
    axs[i].set_xlabel("Salary")
    axs[i].set_ylabel("Frequency")

# Adjust layout to avoid overlapping titles
plt.tight_layout()

# Save the entire figure with all subplots as a single .png file
plt.savefig("data/exported/AnnualSalary_Histogram_All_Departments.png")
plt.show()
```

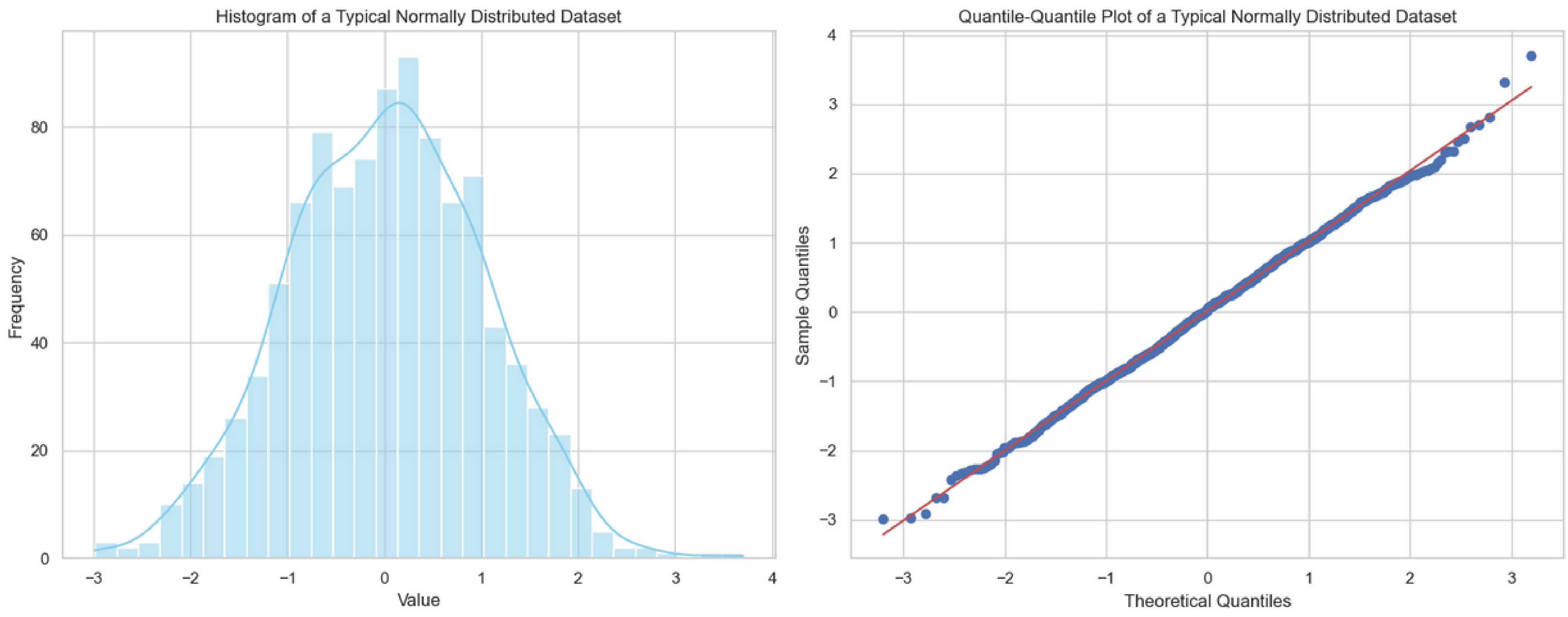
Departmental histogram plots for annual salaried workers can be found within github @ /data/exported under the filename "AnnualSalary_Histogram_All_Departments":
Link: AnnualSalary_Histogram_All_Departments

As observed from the plots, with the exception of Departments AUD, OEM, PHD, RMO, most of the departments are mostly right-skewed & not normally distributed

Hence the significance of z-score value on the selection of the top 5 department with regards to salary disparity is lowered

Thus, emphasis will be placed on CV for departmental salary disparity evaluation as the annual salary of most departments follow a non-normal distribution

Histogram & Q-Q plot of a dataset with normal distribution for Reference



I’ve included both a histogram and a Q-Q plot for a typical normally distributed dataset for reference purposes

Dataset Preparation in SQL [1/2]

Creation of categorical 'Hourly_Annual_Salaried_Employee' column for housekeeping purposes

- Having established that our dataset contains salary information of BOTH hourly and annual waged workers, the SQL query below was written to categorize hourly and annual waged workers for housekeeping purposes

```sql
/*
        SQL Query to create categorical column "Hourly_Annual_Salaried_Employee". For Salary <=10000 we
use "Hourly Salaried Employee" and for  Salary > 10000 we use "Annual Salaried Employee"
*/
-- Run this to drop self-created categorical column
ALTER TABLE EmployeeSalaries_Disparity_Dataset.dbo.Employee_Salaries$
DROP COLUMN IF EXISTS Hourly_Annual_Salaried_Employee;

-- Adds an empty column name called 'Hourly_Annual_Salaried_Employee'
ALTER TABLE EmployeeSalaries_Disparity_Dataset.dbo.Employee_Salaries$
ADD Hourly_Annual_Salaried_Employee VARCHAR(50);

-- Updates 'Hourly_Annual_Salaried_Employee' column with 'Hourly Salaried Employee' & 'Annual Salaried
Employee' based on salary
UPDATE EmployeeSalaries_Disparity_Dataset.dbo.Employee_Salaries$
SET Hourly_Annual_Salaried_Employee =
    CASE
        WHEN Salary <= 10000 THEN 'Hourly Salaried Employee'
        WHEN Salary > 10000 THEN 'Annual Salaried Employee'
    END;
```

- Following which, SQL queries were iteratively built upon to obtain the final SQL query used for departmental analysis

Dataset Preparation in SQL [2/2]

Creation of final query used for analysis

The final query used for departmental salary disparity analysis contains the columns listed below:

1. Standard Deviation
2. Average Salary
3. Coefficient Of Variation
4. Outlier Count based off Z-Score values

[Github Link to SQL Codeblock]

```sql
/*
        FINAL QUERY USED FOR ANALYSIS: Dept Std Dev, Avg Salary, CV, Outlier Count based off Z-Score values
*/
WITH DepartmentStats AS
(SELECT Department,
       STDEV(salary) AS Dept_Std_Dev_Salary,
       AVG(salary) AS Dept_Avg_Salary
FROM EmployeeSalaries_Disparity_Dataset.dbo.Employee_Salaries$
WHERE Salary > 10000
GROUP BY Department
),
DepartmentOutliers AS (
SELECT emp.Department, emp.Salary, ds.Dept_Std_Dev_Salary, ds.Dept_Avg_Salary,
        (emp.Salary - ds.Dept_Avg_Salary)/ds.Dept_Std_Dev_Salary AS Z_Score
FROM EmployeeSalaries_Disparity_Dataset.dbo.Employee_Salaries$ AS emp
INNER JOIN DepartmentStats AS ds ON emp.Department = ds.Department
WHERE emp.Salary > 10000
)
SELECT ds.Department,
       ROUND(ds.Dept_Std_Dev_Salary,2) AS Dept_Std_Dev_Salary,
       ROUND(ds.Dept_Avg_Salary,2) AS Dept_Avg_Salary,
       ROUND((ds.Dept_Std_Dev_Salary / ds.Dept_Avg_Salary),2)*100 AS CoefficientOfVariation, -- % Coefficient of variation measures the relative variability of data; lower values indicate less variability, while higher values indicate more variability.
       SUM(CASE WHEN (do.Z_Score > 1.96 OR do.Z_Score < -1.96) THEN 1 ELSE 0 END) AS Outlier_Count -- Tweakable Z_Score Threshold Values
FROM DepartmentStats AS ds
LEFT JOIN DepartmentOutliers AS do ON ds.Department = do.Department
GROUP BY ds.Department, ds.Dept_Std_Dev_Salary, ds.Dept_Avg_Salary, (ds.Dept_Std_Dev_Salary / ds.Dept_Avg_Salary)
ORDER BY Outlier_Count DESC, CoefficientOfVariation DESC
```

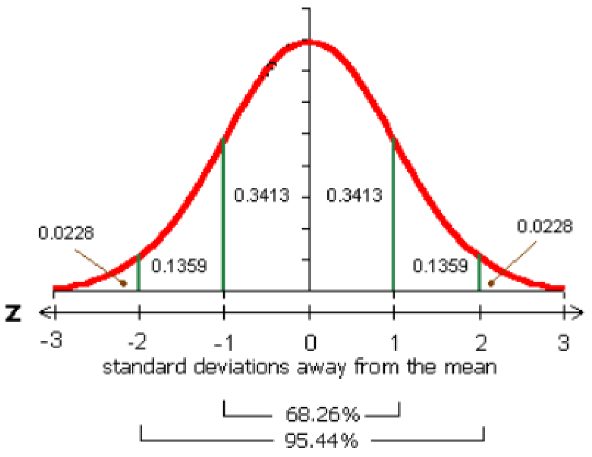
esults | Messages

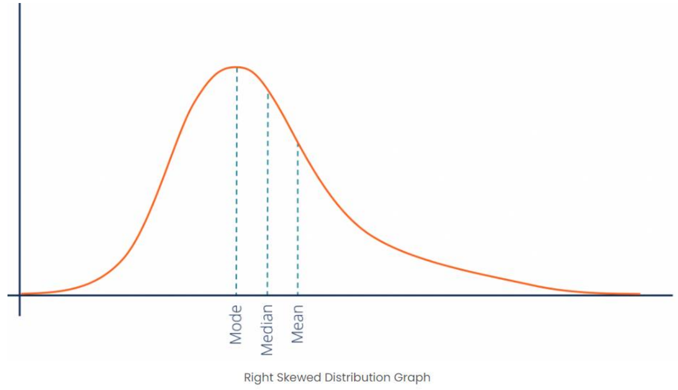
Department	Dept_Std_Dev_Salary	Dept_Avg_Salary	CoefficientOfVariation	Outlier_Count
PWD	22379.82	54081.39	41	47
POL	19520.08	64627.41	30	37
HSD	20283.92	57033.48	36	24
PAR	18245.89	48997.58	37	18
PUD	21055.34	53733.42	39	17
SHF	17918.59	58587.35	31	14
CIT	23292.89	85091.31	27	11
FIR	18007.29	72818.45	25	9
MUS	22527.68	48815.84	46	6
LIB	19056.55	51307.8	37	6
EMS	21898.89	70420.29	31	6
PLN	22612.2	63219.85	36	5

Formula & Significance of calculated columns from SQL Query

Statistical Summary of Annual Salaried Employees [>$10,000]		
	Formula	**Significance & Explanation**
Standard Deviation	$\sigma = \sqrt{\frac{\Sigma(x_i - \mu)^2}{N}}$ σ : population standard deviation N : the size of the population x_i : each value from the population μ : the population mean	The standard deviation of salary within each department reveals the spread of salaries. Higher standard deviation indicates greater disparities. Departments with high standard deviation is one indicator that a department might have salary inequalities.
Average/Mean	$A = \frac{1}{n}\sum_{i=1}^{n} a_i$ A : arithmetic mean N : number of values a_i : data set values	Departments with salaries above the average might be providing better compensation, while those below could indicate potential disparities. In this case, I placed less emphasis on the average of a department as we're interested in analyzing the 'spread'/ salary disparity within each department. Furthermore, it is unfair to compare the average salary of a revenue-generating core department versus a non-core department The Average is instead is used to determine CV, Z-Score, which is eventually used to calculate the count of outliers
Coefficient of Variation [CV]	$CV = \frac{\sigma}{\mu} * 100$ σ: *Department Standard Deviation* μ: *Department Mean/ Average*	The Coefficient of Variation (CV) indicates the size of a standard deviation in relation to its mean. The higher the CV, the greater the dispersion level around the mean, which indicates potential disparities in pay across employees. **CV can be useful in comparing data sets with different units or widely different means, which is the case in this data set. Hence, CV is a larger weighing factor during departmental salary disparity evaluation**
Z-Score	$Z = \frac{x - \mu}{\sigma}$ Z : standard score x : observed value μ : mean of the sample σ: standard deviation of the sample	The Z-Score is a measure of how many standard deviations a data point is away from the mean. The Z-Score threshold used for this analysis is ±1.96, which corresponds to ~ 95% confidence level for a two-tailed test, meaning that about 95% of the data should fall within that range in a **normally distributed dataset**. Therefore, any data point with a Z-Score greater than +1.96 or less than -1.96 is considered an outlier at the 5% significance level. Because of efficacy of using Z-Scores to determine outlier values are somewhat diminished when applied to datasets that are not normally distributed (such as our right-skewed dataset). I've placed lesser weightage on the derived "outlier count" column when determining departments that show the most variance and discrepancy in salary
Outlier Count [based on Z-Score threshold]	N/A	Count of "Outliers" present within each department as defined by the Z-Score Threshold of ± 1.96

Normally Distributed Dataset

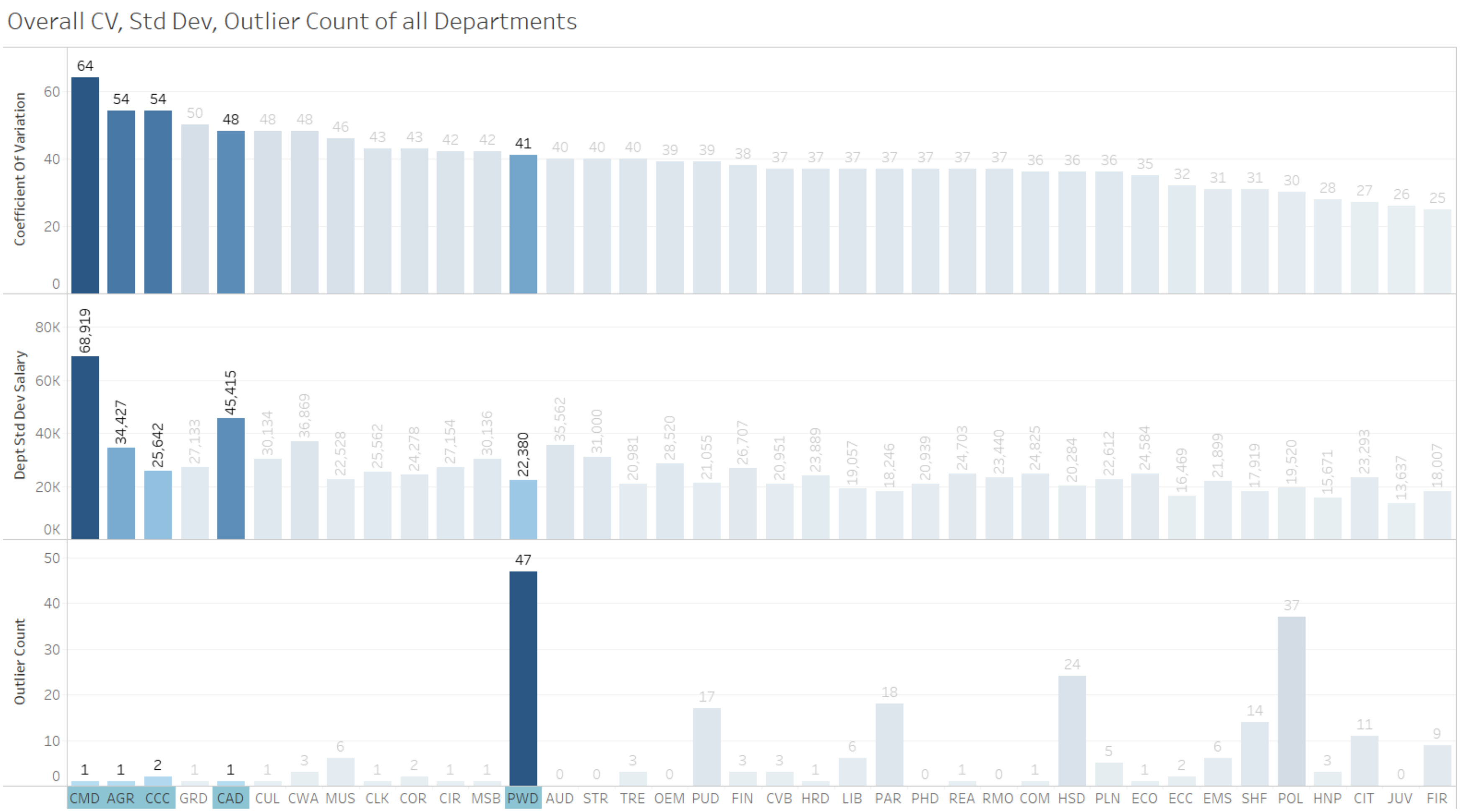




Our right-skewed dataset, where the mean does not correspond to the peak

Overall CV, Std Deviation, Outlier Count of all Departments

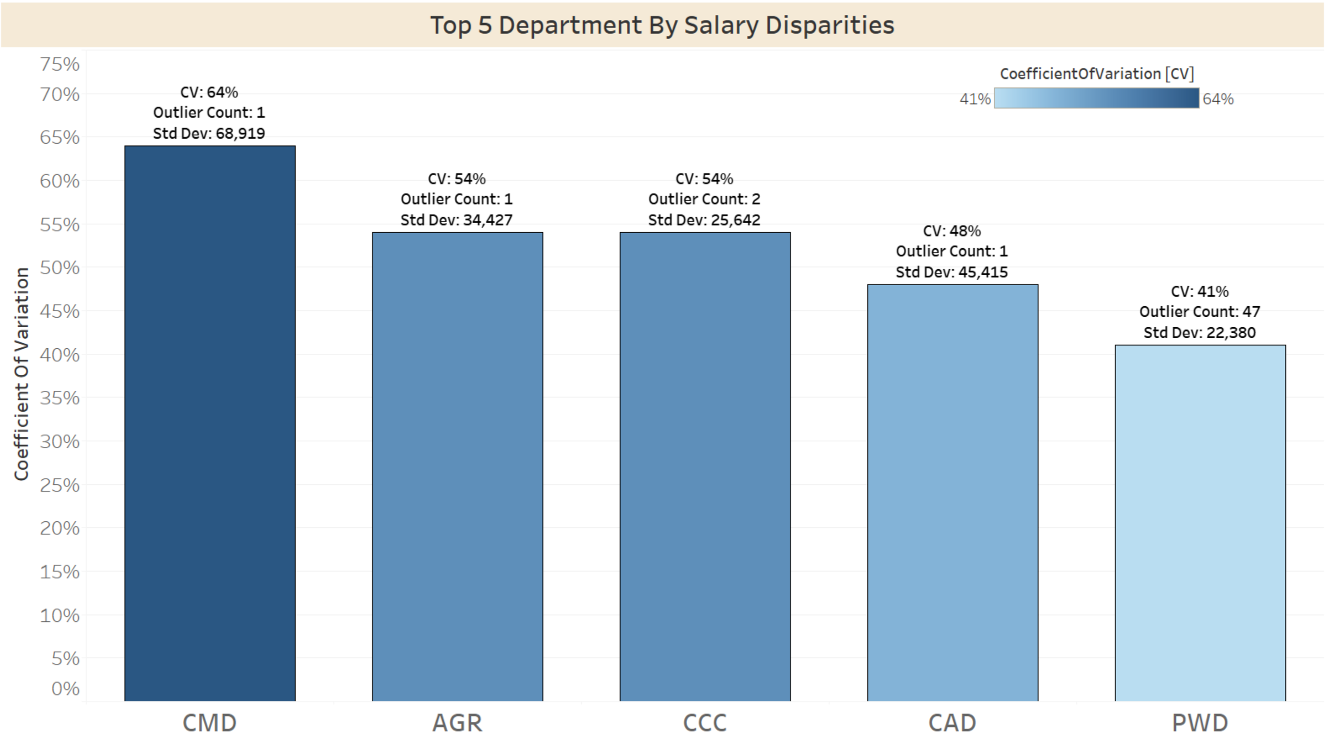
A plot based on statistical metrics obtained from the SQL query (Coefficient of Variation, Standard Deviation & outlier counts) was made below using Tableau, along with the reason for department selection:



Department	Reason for selection
CMD	Highest CV & Std Dev, 1 outlier
AGR	2nd Highest CV, same as CCC department. However, it has higher standard deviation as compared to CCC department. Also has Outlier count of 1
CCC	Quite similar to AGR in terms of CV and Outlier Count, but has less Standard Deviation than AGR, which indicates less salary 'spread' or variance amongst the department
CAD	Slightly lower CV compared to GRD, but has higher standard deviation compared to GRD. Lower CV as compared to AGR and CCC. Has the highest Standard Deviation compared to it's peers with same CV value of 48
PWD	Highest Outlier count in the dataset, fairly similar CV compared to CAD department. POL department which had the 2nd highest Outlier count was not selected as it had significantly lower CV & Standard deviation value as compared to PWD

Findings & Recommendation

The plot below summarizes the top 5 departments that have been selected for management to review, with regards to having the most variance and discrepancy in salary



Conclusion:
PWD Department being flagged as having a high amount of salary spread is validated as it had the highest outlier count and has a moderately high CV value. However, Management should also look into the other departments listed in the plot above for salary discrepancy review

END OF PRESENTATION