Missouri State Employee Overview

DataChamps

April 19, 2024

Here is the GitHub URL: Github Repo

Abstract

Our goal was to analyze the 2021 Missouri State Employee report. After examining the data set we came up with number of goals we'd want to achieve from a deeper break down of our data set, those being(Total spent on employee salary, Average employee salary, How many employees for each agency, Total income for each agency, Average salary for each agency,and Highest and lowest paid employee). In this report we will go over how we use Pyspark in Jupyterlab, and the steps we took to accomplish those goals.

1 Implementation steps

1.1 Imports and Preporation:

```
from pyspark.sql import SparkSession
spark = SparkSession.builder.getOrCreate()
from pyspark.sql.functions import round
from pyspark.sql.types import DecimalType, DoubleType
from pyspark.sql.functions import desc, asc, exp, max, col
```

```
from datetime import datetime, date
import pandas as pd
from pyspark.sql import Row
df = spark.createDataFrame([
    Row(a=1, b='string1', c='string1', d='string1', e=2.00),
    Row(a=2, b='string2', c='string2', d='string2', e=3.00),
    Row(a=4, b='string3', c='string3', d='string3', e=5.00)
])
df
```

```
DataFrame[a: bigint, b: string, c: string, d: string, e: double]
```

```
sales = spark.read.format('csv').option('header','true').load('2021_State_Employee_Pay.csv')
sales.createOrReplaceTempView('em')
sales.show(truncate=False)
```

Figure 1: Implementing Data set and prep

To start our data manipulation we have to start with our imports and set up. From pyspark.sql was the header for most of our imports and tools use. Adding types and functions to our header allowed for us to use tools like round, DecimalType, DoubleType, desc, asc, exp, max, col.

1.2 Data Formatting:

Before reading in we had to format our data frame to properly be able to handle our cvs file. Our data set has columns being Calendar Year - integertype, Agency Name - stringtype, Position Title - stringtype, Employee Name - stringtype, and YTD Gross Pay - doubletype. Using df = spark.createDataFrame we were able to pre set the data types we wanted each column to be so that we could manipulate our data set to achieve the goals we set.

```
from datetime import datetime, date
import pandas as pd
from pyspark.sql import Row
df = spark.createDataFrame([
    Row(a=1, b='string1', c='string1', d='string1', e=2.00),
    Row(a=2, b='string2', c='string2', d='string2', e=3.00),
    Row(a=4, b='string3', c='string3', d='string3', e=5.00)
])
df
```

```
DataFrame[a: bigint, b: string, c: string, d: string, e: double]
```

Figure 2: Data Formatting

1.3 Reading in Data set:

```
sales = spark.read.format('csv').option('header','true').load('2021_State_Employee_Pay.csv')
sales.createOrReplaceTempView('em')
sales.show(truncate=False)
```

```
+-------------+-----------+----------------------+------------------------+-------------+
|Calendar_Year|Agency_Name|Position_Title        |Employee_Name           |YTD_Gross_Pay|
+-------------+-----------+----------------------+------------------------+-------------+
|2021         |AGRICULTURE|ACCOUNTANT            |KLEINDIENST, ANGELA F   |44054.5      |
|2021         |AGRICULTURE|ACCOUNTANT            |WOOD, KAREN M.          |39339.97     |
|2021         |AGRICULTURE|ACCOUNTS SUPERVISOR   |WALKER, JOE E.          |53821.2      |
|2021         |AGRICULTURE|ADMIN SUPPORT ASSISTANT|BICKERTON, HAILEY ANN   |29484        |
|2021         |AGRICULTURE|ADMIN SUPPORT ASSISTANT|BIRDWELL, RHIANNON      |15340.08     |
|2021         |AGRICULTURE|ADMIN SUPPORT ASSISTANT|HALL, STACY A.          |20208.3      |
|2021         |AGRICULTURE|ADMIN SUPPORT ASSISTANT|HENRY, PAMELA A         |14742        |
|2021         |AGRICULTURE|ADMIN SUPPORT ASSISTANT|JONES, MEGAN L          |15541.71     |
|2021         |AGRICULTURE|ADMIN SUPPORT ASSISTANT|KIRSCH, NICOLE LEANNE   |31167.32     |
|2021         |AGRICULTURE|ADMIN SUPPORT ASSISTANT|LAWSON, TRACY L.        |9544.74      |
```

Figure 3: Reading data and Sample

Once we are done formatting we can read in our data set and create a temporary view. Using read and format commands we are able to import our data with header attached and store it in a data frame. We create a temp view so that we don't alter our original data as we answer our goals. Lastly we have to check too see if our data is read in properly, using sales.show(truncate=False) we are able to see the full content of our data frame and start working on our goals.

2 Goals

2.1 Calculate the total amount spent on employee salary.

To compute the overall expenditure on employee wages, we aggregate "YTD_Gross_Pay" data by applying the sum operation using the agg() function. We then format the resulting total using the DecimalType to ensure accuracy. This aggregated sum serves as a crucial metric for various purposes, such as budgeting, annual growth analysis, and benchmarking against other states' budgets. It offers valuable insights into how funds are allocated for employee compensation within Missouri's state government, aiding in decision-making and resource distribution. The total amount expended on

salaries during the specified period, amounting to $2,160,391,753.17, indicates a significant volume of financial transactions, underscores the substantial financial commitment towards supporting the state workforce.

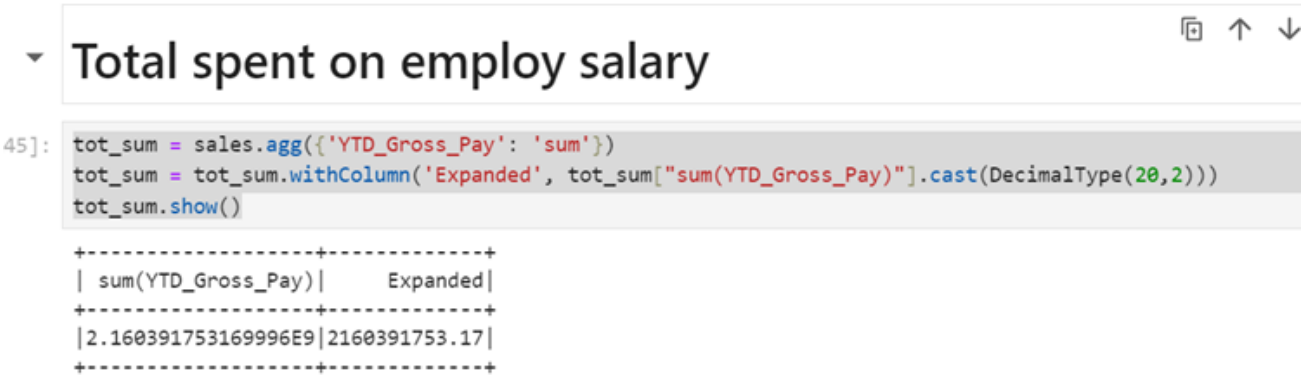

Figure 4: Total amount spent on employee salary

2.2 Calculate the average amount spent on the employee salary.

To derive the average salary per employee, we utilize PySpark's agg() function, aggregating the "YTD Gross Pay" column with the 'avg' method. This computation yields the mean salary across all employees. The result is then converted to DecimalType to ensure accuracy and displayed using the show() function. Understanding the average salary facilitates comparisons among individual salaries and enhances the attraction of potential hires during recruitment. For example, it offers insights into an employee's salary positioning within the organization. In the provided outcome, the average yearly gross pay stands at $30,628.65. This metric acts as a yardstick for assessing salary competitiveness and informs decisions regarding compensation strategies and recruitment initiatives.

Figure 5: Average amount spent on the employee salary

2.3 Calculate the number of employees for each department.

To ascertain the employee count for each agency, we employ PySpark groupBy() function to organize the dataset by "Agency Name" and then utilize the count() function to aggregate the counts. Sorting the outcomes by "Agency Name" furnishes a systematic overview. The resulting data showcases Corrections as the agency with the highest employee count, boasting 12,530 employees. Conversely, OOLG emerges as the agency with the lowest count, employing only 19 individuals. This data sheds light on the distribution of the workforce across various agencies within Missouri's state government, highlighting notable differences in size. Such insights potentially indicate differing staffing needs and

Number of employess

```
[156]: agen_employ = sales.groupBy("Agency_Name").agg({"Agency_Name": "count"}).sort("Agency_Name")
       agen_employ.sort(desc("count(Agency_Name)")).show(24, truncate=False)
```

```
+---------------------------------+-------------------+
|Agency_Name                      |count(Agency_Name)|
+---------------------------------+-------------------+
|CORRECTIONS                      |12530              |
|MENTAL HEALTH                    |10089              |
|SOCIAL SERVICES                  |8362               |
|PUBLIC SAFETY                    |7385               |
|TRANSPORTATION                   |7170               |
|JUDICIARY                        |5017               |
|ELEMENTARY AND SECONDARY EDUCATION|2845               |
|OFFICE OF ADMINISTRATION         |2419               |
|HEALTH AND SENIOR SERVICES       |2311               |
|NATURAL RESOURCES                |2310               |
|CONSERVATION                     |2303               |
|REVENUE                          |1929               |
|AGRICULTURE                      |1273               |
|COMMERCE AND INSURANCE           |1076               |
|LEGISLATURE                      |873                |
|LABOR AND INDUSTRIAL RELATIONS   |824                |
|HIGHER EDUCATION AND WORKFORCE DEV|514                |
|OFFICE OF ATTORNEY GENERAL       |505                |
|OFFICE OF SECRETARY OF STATE     |315                |
|ECONOMIC DEVELOPMENT             |210                |
|OFFICE OF STATE AUDITOR          |161                |
|OFFICE OF STATE TREASURER        |56                 |
|OFFICE OF GOVERNOR               |39                 |
|OFFICE OF LIEUTENANT GOVERNOR    |19                 |
+---------------------------------+-------------------+
```

Figure 6: Number of employees for each department.

organizational structures across agencies. Armed with this knowledge, stakeholders can make informed decisions regarding resource allocation and devise workforce management strategies aimed at bolstering operational efficiency.

2.4 Calculate the total income for each department

Total Salary per agency

```
[171]: agen_tot = sales.groupBy("Agency_Name").agg({"YTD_Gross_Pay": "sum"}).sort("Agency_Name")
       agen_tot = agen_tot.withColumn('Total_per_agency', agen_tot["sum(YTD_Gross_Pay)"].cast(DecimalType(20,2)))
       agen_tot.sort(desc("sum(YTD_Gross_Pay)")).show(24, truncate=False)
```

```
+---------------------------------+--------------------+----------------+
|Agency_Name                      |sum(YTD_Gross_Pay)  |Total_per_agency|
+---------------------------------+--------------------+----------------+
|CORRECTIONS                      |3.822000074700007E8 |382200007.47    |
|MENTAL HEALTH                    |2.7705459687000114E8|277054596.87    |
|PUBLIC SAFETY                    |2.3506643366000026E8|235066433.66    |
|SOCIAL SERVICES                  |2.3418771013999888E8|234187710.14    |
|TRANSPORTATION                   |2.326571463500011E8 |232657146.35    |
|JUDICIARY                        |2.008815517899995E8 |200881551.79    |
|OFFICE OF ADMINISTRATION         |8.844222656999959E7 |88442226.57     |
|HEALTH AND SENIOR SERVICES       |7.883300668999992E7 |78833006.69     |
|CONSERVATION                     |6.992013162000003E7 |69920131.62     |
|ELEMENTARY AND SECONDARY EDUCATION|6.925081386999997E7 |69250813.87     |
|NATURAL RESOURCES                |6.640370936000007E7 |66403709.36     |
|REVENUE                          |4.607970200000001E7 |46079702.00     |
|COMMERCE AND INSURANCE           |4.1321026110000014E7|41321026.11     |
|LEGISLATURE                      |2.969761229000006E7 |29697612.29     |
|LABOR AND INDUSTRIAL RELATIONS   |2.9692478259999957E7|29692478.26     |
|OFFICE OF ATTORNEY GENERAL       |1.9561059030000005E7|19561059.03     |
|AGRICULTURE                      |1.710069344E7       |17100693.44     |
|HIGHER EDUCATION AND WORKFORCE DEV|1.4353807269999988E7|14353807.27     |
|OFFICE OF SECRETARY OF STATE     |9082329.97          |9082329.97      |
|ECONOMIC DEVELOPMENT             |6982084.72          |6982084.72      |
|OFFICE OF STATE AUDITOR          |6426858.909999998   |6426858.91      |
|OFFICE OF STATE TREASURER        |2280548.1200000006  |2280548.12      |
|OFFICE OF GOVERNOR               |2030056.8900000004  |2030056.89      |
|OFFICE OF LIEUTENANT GOVERNOR    |886161.7699999999   |886161.77       |
+---------------------------------+--------------------+----------------+
```

Figure 7: Total income for each department

To compute the total income per agency, the Missouri state employee dataset undergoes grouping by the "Agency Name" column, followed by aggregation of the "YTD Gross Pay" column's sum. Sorting the results by agency name unveils the total income for each entity. Notably, Corrections stands out with the highest total income of $382,200,007.47, underscoring its significant financial presence within

the state employee payroll. Conversely, the Office of Lieutenant Governor (OOLG) registers the lowest total income at $886,161.77, indicating its relatively minor financial impact within the dataset. This analysis provides valuable insights into the financial distribution among different agencies, spotlighting those making substantial contributions to the overall income pool and those with more modest financial footprints.

2.5 Calculate the average salary for each department.

To determine the average salary per agency, I utilized PySpark groupBy function to group the data by "Agency Name" and applied the avg function to calculate the average of the "YTD Gross Pay" column. Sorting the results by agency name, I refined the average values by casting them to a DecimalType with precision 20 and scale 2 for enhanced accuracy. Sorting the outcome in descending order by average salary, I found the Office of Governor (OOG) to have the highest average salary at $52,052.74, while the Agriculture department (Ag) had the lowest at $13,433.38. The state-wide average salary across all agencies was calculated at $30,628.65. This analysis underscores substantial variations in average salaries among different agencies, with some significantly exceeding and others falling below the state average. Such insights are pivotal for grasping the financial dynamics within the state employment framework and can serve as valuable inputs for employees and policymakers alike in decision-making processes.

Avg Agency salary

```
[172]: agen_avg = sales.groupBy("Agency_Name").agg({"YTD_Gross_Pay": "avg"}).sort("Agency_Name")
       agen_avg = agen_avg.withColumn('Avg_per_agency', agen_avg["avg(YTD_Gross_Pay)"].cast(DecimalType(20,2)))
       agen_avg.sort(desc("Avg_per_agency")).show(24, truncate=False)
```

```
+---------------------------------+-----------------+--------------+
|Agency_Name                      |avg(YTD_Gross_Pay)|Avg_per_agency|
+---------------------------------+-----------------+--------------+
|OFFICE OF GOVERNOR               |52052.740769230775|52052.74      |
|OFFICE OF LIEUTENANT GOVERNOR    |46640.093157894735|46640.09      |
|OFFICE OF STATE TREASURER        |40724.073571428584|40724.07      |
|JUDICIARY                        |40040.17376719145 |40040.17      |
|OFFICE OF STATE AUDITOR          |39918.378322981356|39918.38      |
|OFFICE OF ATTORNEY GENERAL       |38734.770356435656|38734.77      |
|COMMERCE AND INSURANCE           |38402.44062267659 |38402.44      |
|OFFICE OF ADMINISTRATION         |36561.482666390904|36561.48      |
|LABOR AND INDUSTRIAL RELATIONS   |36034.56099514558 |36034.56      |
|HEALTH AND SENIOR SERVICES       |34112.07559065337 |34112.08      |
|LEGISLATURE                      |34017.883493699956|34017.88      |
|ECONOMIC DEVELOPMENT             |33248.022476190476|33248.02      |
|TRANSPORTATION                   |32448.6954463042  |32448.70      |
|PUBLIC SAFETY                    |31830.2550656737  |31830.26      |
|CORRECTIONS                      |30502.793892258633|30502.79      |
|CONSERVATION                     |30360.456630481996|30360.46      |
|OFFICE OF SECRETARY OF STATE     |28832.793555555556|28832.79      |
|NATURAL RESOURCES                |28746.194528138556|28746.19      |
|SOCIAL SERVICES                  |28006.183944032393|28006.18      |
|HIGHER EDUCATION AND WORKFORCE DEV|27925.69507782099 |27925.70      |
|MENTAL HEALTH                    |27461.05628605423 |27461.06      |
|ELEMENTARY AND SECONDARY EDUCATION|24341.235103690677|24341.24      |
|REVENUE                          |23887.87039917056 |23887.87      |
|AGRICULTURE                      |13433.38054988217 |13433.38      |
+---------------------------------+-----------------+--------------+
```

Figure 8: Average salary for each department

2.6 Calculate the highest and lowest paid employee for each department.

To determine the highest and lowest-paid employees, we'll analyze the salary data for 2021 Missouri state employees using PySpark. By sorting the dataset based on salary, we can easily identify the employee with the highest salary, likely a psychiatrist or another high-ranking official. Conversely, the lowest-paid employee would be found at the bottom of the sorted list, possibly an entry-level or

part-time position within one of the agencies. This analysis provides valuable insights into the salary distribution across different roles within the state government, helping individuals understand the range of earnings potential. Moreover, it offers transparency regarding the financial rewards associated with various positions, aiding in career decision-making and resource allocation within state agencies.

agency total highest to lowest

[162]:
```
highest_em = agen_tot.sort(desc("sum(YTD_Gross_Pay)"))
highest_em.show(24, truncate=False)
```

```
+---------------------------------+-------------------+----------------+
|Agency_Name                      |sum(YTD_Gross_Pay) |Total_per_agency|
+---------------------------------+-------------------+----------------+
|CORRECTIONS                      |3.82200007470007E8 |382200007.47    |
|MENTAL HEALTH                    |2.7705459687000114E8|277054596.87    |
|PUBLIC SAFETY                    |2.3506643366000026E8|235066433.66    |
|SOCIAL SERVICES                  |2.3418771013999888E8|234187710.14    |
|TRANSPORTATION                   |2.3265714635000011E8 |232657146.35    |
|JUDICIARY                        |2.008815517899995E8 |200881551.79    |
|OFFICE OF ADMINISTRATION         |8.8442226569999959E7 |88442226.57     |
|HEALTH AND SENIOR SERVICES       |7.883300668999992E7 |78833006.69     |
|CONSERVATION                     |6.992013162000003E7 |69920131.62     |
|ELEMENTARY AND SECONDARY EDUCATION|6.925081386999997E7 |69250813.87     |
|NATURAL RESOURCES                |6.640370936000007E7 |66403709.36     |
|REVENUE                          |4.6079702000000001E7 |46079702.00     |
|COMMERCE AND INSURANCE           |4.1321026110000014E7|41321026.11     |
|LEGISLATURE                      |2.969761229000006E7 |29697612.29     |
|LABOR AND INDUSTRIAL RELATIONS   |2.9692478259999957E7|29692478.26     |
|OFFICE OF ATTORNEY GENERAL       |1.9561059030000005E7|19561059.03     |
|AGRICULTURE                      |1.710069344E7       |17100693.44     |
|HIGHER EDUCATION AND WORKFORCE DEV|1.4353807269999988E7|14353807.27     |
|OFFICE OF SECRETARY OF STATE     |9082329.97          |9082329.97      |
|ECONOMIC DEVELOPMENT             |6982084.72          |6982084.72      |
|OFFICE OF STATE AUDITOR          |6426858.909999998   |6426858.91      |
|OFFICE OF STATE TREASURER        |2280548.1200000006  |2280548.12      |
|OFFICE OF GOVERNOR               |2030056.8900000004  |2030056.89      |
|OFFICE OF LIEUTENANT GOVERNOR    |886161.7699999999   |886161.77       |
+---------------------------------+-------------------+----------------+
```

Figure 9: Highest and lowest paid employee for each department

3 Conclusion

In our analysis of 2021 Missouri state employee pay, we found compelling patterns that shed light on the dynamics within different agencies. Firstly, it's evident that the Office of Governor, Office of Lieutenant Governor, and Office of State Treasurer stand out with the highest average salaries, despite having fewer employees and lower total income. This suggests that while they cost less to employ, they offer significant value, making them desirable workplaces, especially for those seeking smaller units with higher median salaries. On the other hand, Corrections, with its large workforce, poses challenges in terms of space and financial resources for operations. Despite being a heavy burden on total income, its position in the middle for average pay makes it less appealing as a workspace, albeit offering ample opportunities for employment.

Furthermore, our analysis highlights the prominence of psychiatrists as the top earners among state employees. This finding underscores the potential for substantial financial rewards in pursuing this profession within the state system. Moreover, the revelation of at least 100 employees indebted to the state underscores an important financial aspect that warrants attention and potentially requires measures to address. Overall, these insights provide valuable guidance for both job seekers and policymakers, informing decisions regarding career paths, resource allocation, and financial management within state agencies.

4 References

- https://spark.apache.org/docs/latest/api/python/index.html
- https://data.mo.gov/dataset/2021-State-Employee-Pay/7j8x-y8ki/about$_data$