DATA ANALYSIS REPORT

Dataset related to research and workflow management

Following is the representation of columns with description

- ProfileID: A unique identifier for each profile.

Dates:

- Created Date: The date when the profile was created.
- Secondary Completed Date: The date when secondary research was completed.
- Close Date: The date when the profile was closed.

Researcher Information:

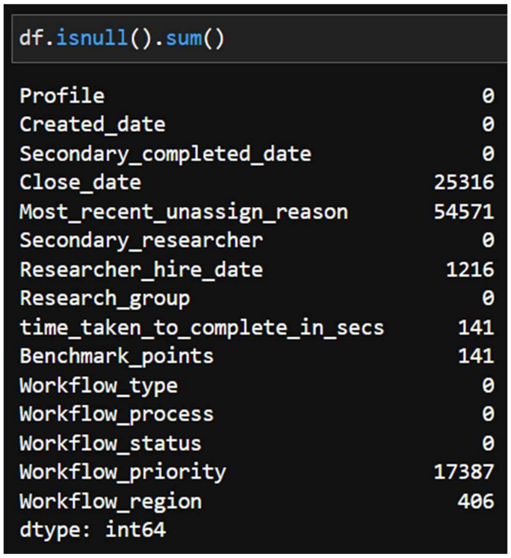
- Most Recent Unassign Reason: The reason for the most recent unassignment.
- Secondary Researcher: The researcher assigned to secondary research.
- Researcher Hire Date: The date when the researcher was hired.
- Research Group: The group to which the researcher belongs.

Performance Metrics:

- Time taken to Complete (secs): The time taken to complete the workflow in seconds.
- Benchmark Points: Points associated with the profile's benchmark performance.

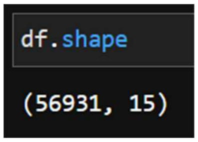
Workflow Details:

- Workflow Type: Type of workflow (e.g., "Co VC - New Round," "Inv VC - New Investor").
- Workflow Process: Specific process within the workflow (e.g., "Survey," "Check In," "Pend Survey").
- Workflow Status: Current status of the workflow (e.g., "Closed," "Pend Survey").
- Workflow Priority: Priority level of the workflow (e.g., "Regular," "High").
- Workflow Region: Geographical region associated with the workflow (e.g., "South America," "North America," "Asia").



```
df.isnull().sum()

Profile                             0
Created_date                        0
Secondary_completed_date            0
Close_date                      25316
Most_recent_unassign_reason     54571
Secondary_researcher                0
Researcher_hire_date             1216
Research_group                      0
time_taken_to_complete_in_secs    141
Benchmark_points                  141
Workflow_type                       0
Workflow_process                    0
Workflow_status                     0
Workflow_priority               17387
Workflow_region                   406
dtype: int64
```

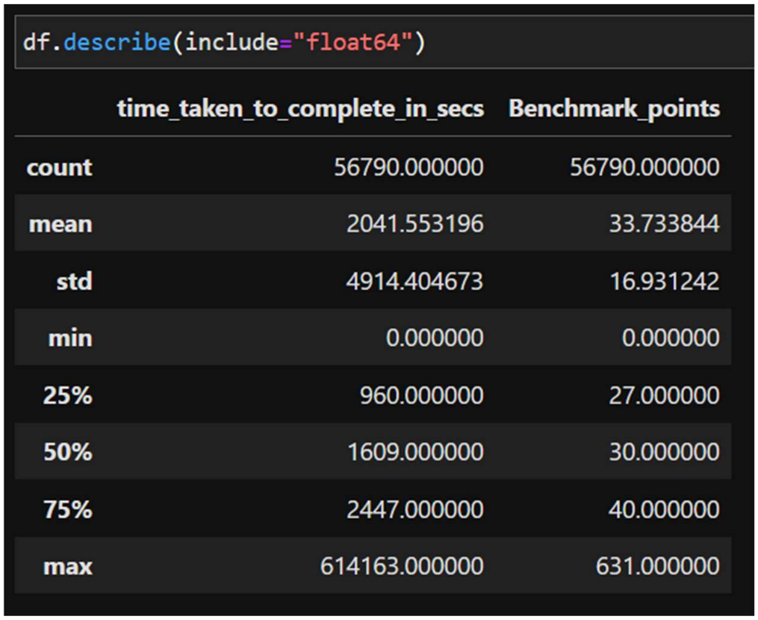
The figure besides provides a summary of the null values present in each column of the given dataset. It's crucial to emphasize that no null values are being filled, particularly in columns such as Dates and Workflow Region. These columns contain factual values, and attempting to predict or fill nulls could lead to inaccurate results



```
df.shape
```

```
(56931, 15)
```

The dataset comprises a total of 56,931 records and includes 15 columns or fields.



```
df.describe(include="float64")
```

	time_taken_to_complete_in_secs	Benchmark_points
count	56790.000000	56790.000000
mean	2041.553196	33.733844
std	4914.404673	16.931242
min	0.000000	0.000000
25%	960.000000	27.000000
50%	1609.000000	30.000000
75%	2447.000000	40.000000
max	614163.000000	631.000000

The provided statistics describe the columns time taken to complete in secs and Benchmark Points. Here's an interpretation of each statistic:

count: The number of non-null entries in the dataset for both columns. In this case, there are 56,790 entries.

mean: The average value of the data. For Time taken to complete the Workflow process (in secs), the average completion time is approximately 2041.55 seconds, and for Benchmark Points, the average is about 33.73.

std: The standard deviation measures the amount of variation or dispersion in the dataset. For Time taken to complete the Workflow process (in secs), the standard deviation is approximately 4914.40, and for Benchmark Points, it's about 16.93.

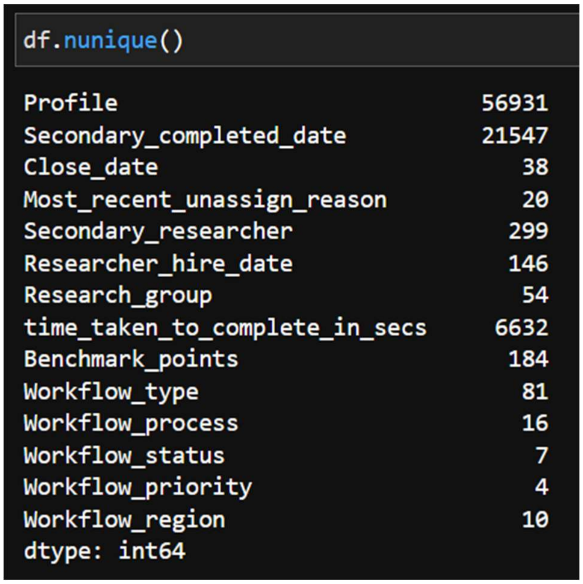
min: The minimum value in the dataset. For Time taken to complete the Workflow process (in secs), the minimum completion time is 0 seconds, and for Benchmark Points, the minimum value is 0.

25% (Q1): The first quartile or the 25th percentile. This is the value below which 25% of the data falls. For Time taken to complete the Workflow process (in secs), 25% of the data has completion times less than or equal to 960 seconds, and for Benchmark Points, 25% of the data has values less than or equal to 27.

50% (Q2): The second quartile or the median. This is the middle value of the dataset. For Time taken to complete the Workflow process (in secs), the median completion time is 1609 seconds, and for Benchmark Points, the median value is 30.

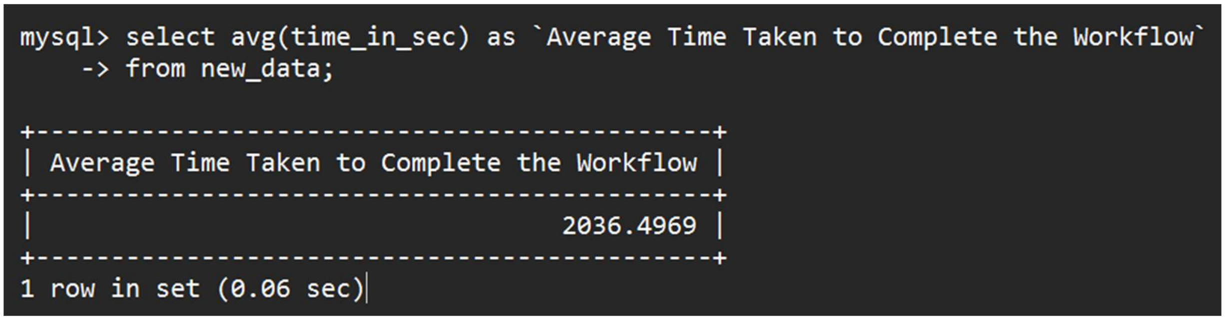
75% (Q3): The third quartile or the 75th percentile. This is the value below which 75% of the data falls. For Time taken to complete the Workflow process (in secs), 75% of the data has completion times less than or equal to 2447 seconds, and for Benchmark Points, 75% of the data has values less than or equal to 40.

max: The maximum value in the dataset. For Time taken to complete the Workflow process (in secs), the maximum completion time is 614163 seconds, and for Benchmark Points, the maximum value is 631.



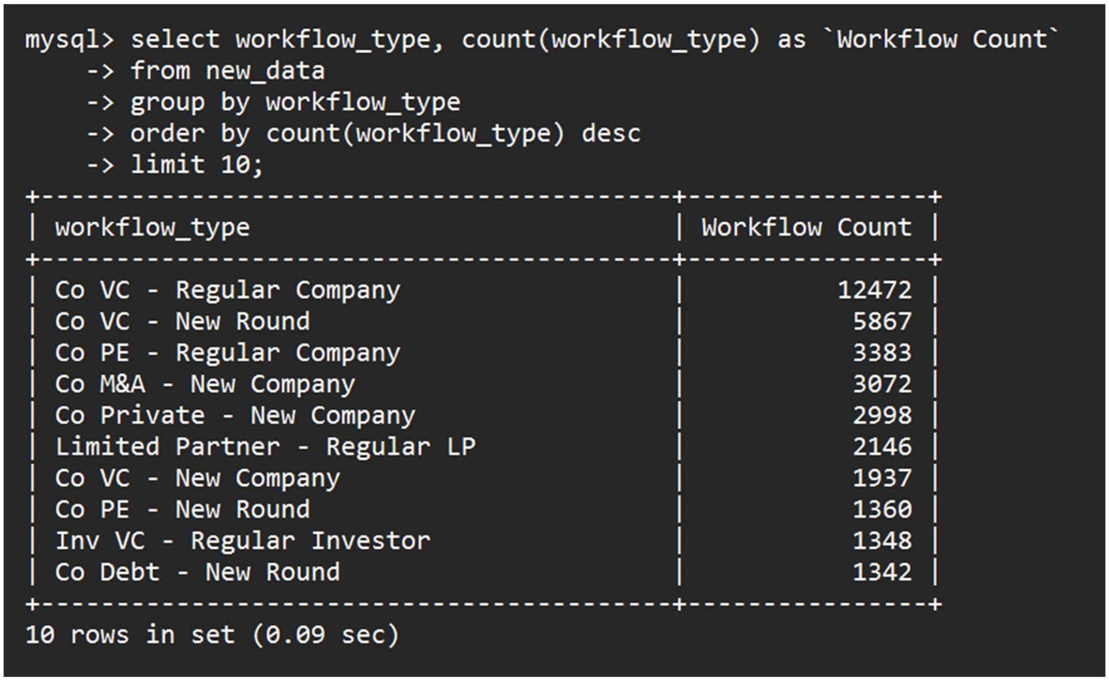
```
df.nunique()

Profile                          56931
Secondary_completed_date         21547
Close_date                          38
Most_recent_unassign_reason         20
Secondary_researcher               299
Researcher_hire_date               146
Research_group                      54
time_taken_to_complete_in_secs    6632
Benchmark_points                   184
Workflow_type                       81
Workflow_process                    16
Workflow_status                      7
Workflow_priority                    4
Workflow_region                     10
dtype: int64
```

The figure illustrates the count of distinct or unique values in each column.



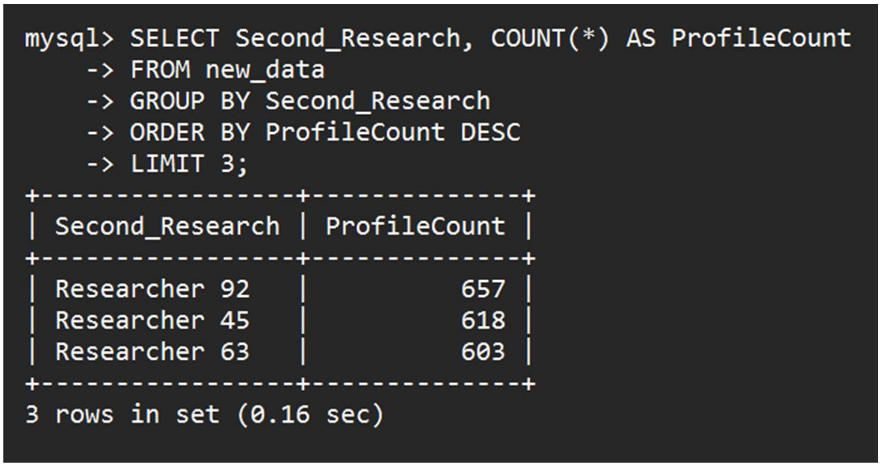
```
mysql> select avg(time_in_sec) as `Average Time Taken to Complete the Workflow`
    -> from new_data;

+--------------------------------------------+
| Average Time Taken to Complete the Workflow |
+--------------------------------------------+
|                                  2036.4969 |
+--------------------------------------------+
1 row in set (0.06 sec)
```

Above SQL query shows the Average time taken to complete the Workflow Process which is 2036.4969 secs.



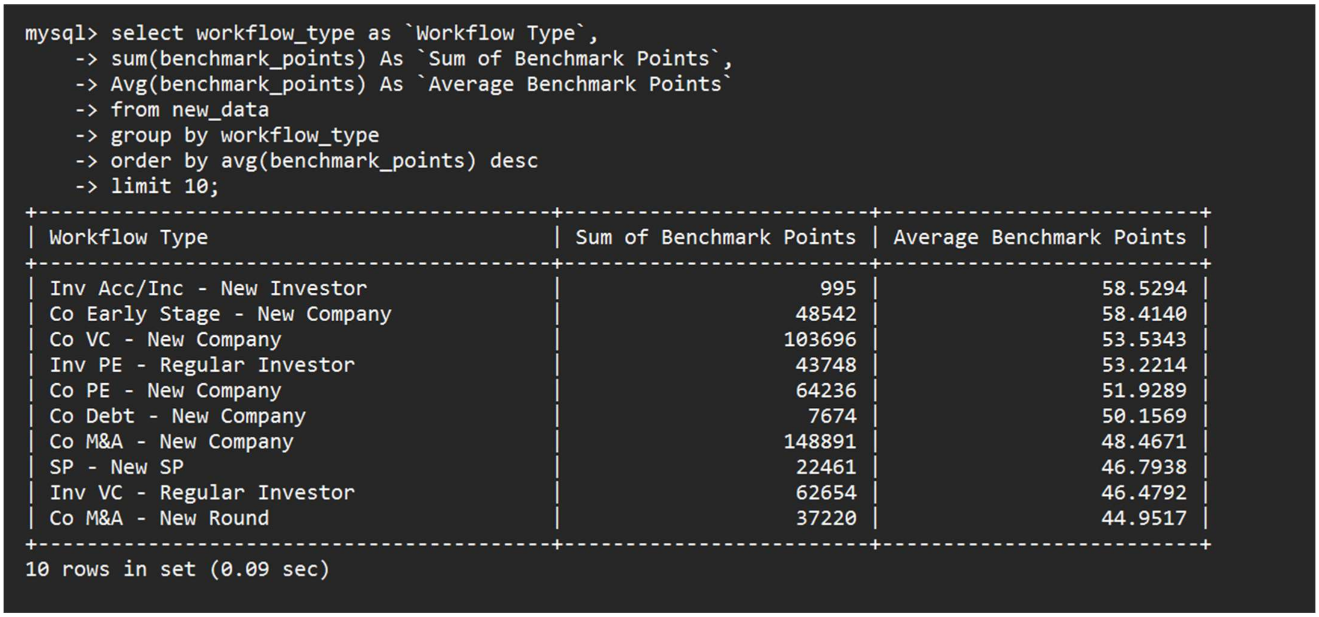
```
mysql> select workflow_type, count(workflow_type) as `Workflow Count`
    -> from new_data
    -> group by workflow_type
    -> order by count(workflow_type) desc
    -> limit 10;
+------------------------------------------+----------------+
| workflow_type                            | Workflow Count |
+------------------------------------------+----------------+
| Co VC - Regular Company                  |          12472 |
| Co VC - New Round                        |           5867 |
| Co PE - Regular Company                  |           3383 |
| Co M&A - New Company                     |           3072 |
| Co Private - New Company                 |           2998 |
| Limited Partner - Regular LP             |           2146 |
| Co VC - New Company                      |           1937 |
| Co PE - New Round                        |           1360 |
| Inv VC - Regular Investor                |           1348 |
| Co Debt - New Round                      |           1342 |
+------------------------------------------+----------------+
10 rows in set (0.09 sec)
```

The SQL query above presents the distribution of Workflow types, with the results limited to 10 outputs.



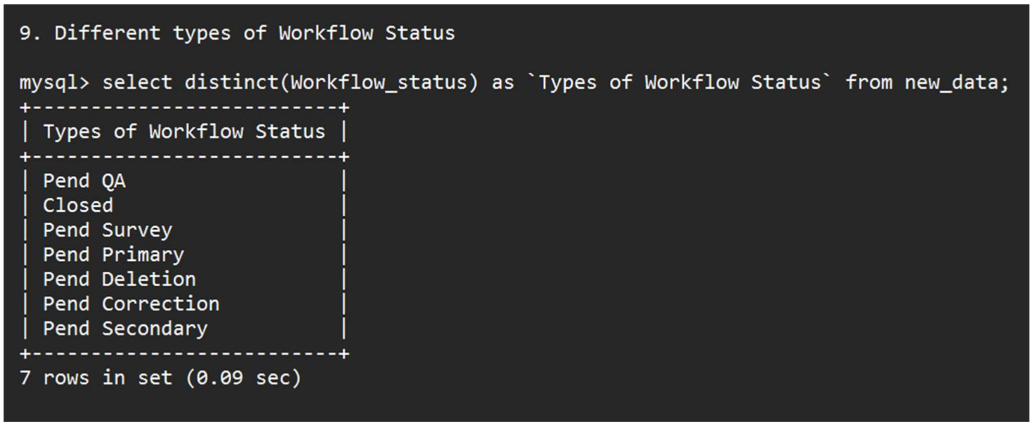
```
mysql> SELECT Second_Research, COUNT(*) AS ProfileCount
    -> FROM new_data
    -> GROUP BY Second_Research
    -> ORDER BY ProfileCount DESC
    -> LIMIT 3;
+-----------------+--------------+
| Second_Research | ProfileCount |
+-----------------+--------------+
| Researcher 92   |          657 |
| Researcher 45   |          618 |
| Researcher 63   |          603 |
+-----------------+--------------+
3 rows in set (0.16 sec)
```

The SQL query above indicates which secondary researcher has been assigned the highest number of profiles. The result is limited to the top 3.



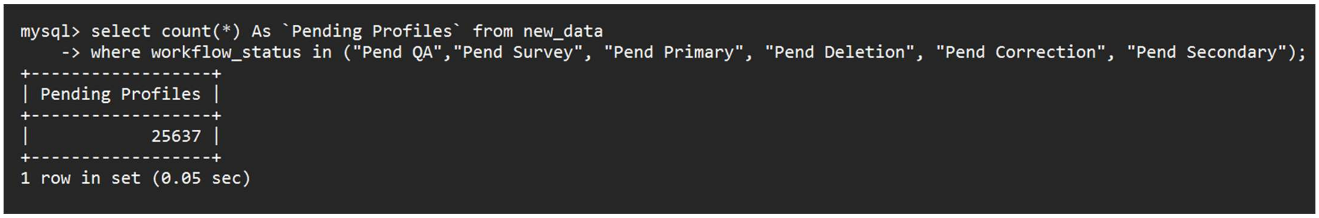
```
mysql> select workflow_type as `Workflow Type`,
    -> sum(benchmark_points) As `Sum of Benchmark Points`,
    -> Avg(benchmark_points) As `Average Benchmark Points`
    -> from new_data
    -> group by workflow_type
    -> order by avg(benchmark_points) desc
    -> limit 10;
+------------------------------------------+-------------------------+--------------------------+
| Workflow Type                            | Sum of Benchmark Points | Average Benchmark Points |
+------------------------------------------+-------------------------+--------------------------+
| Inv Acc/Inc - New Investor               |                     995 |                  58.5294 |
| Co Early Stage - New Company             |                   48542 |                  58.4140 |
| Co VC - New Company                      |                  103696 |                  53.5343 |
| Inv PE - Regular Investor                |                   43748 |                  53.2214 |
| Co PE - New Company                      |                   64236 |                  51.9289 |
| Co Debt - New Company                    |                    7674 |                  50.1569 |
| Co M&A - New Company                     |                  148891 |                  48.4671 |
| SP - New SP                              |                   22461 |                  46.7938 |
| Inv VC - Regular Investor                |                   62654 |                  46.4792 |
| Co M&A - New Round                       |                   37220 |                  44.9517 |
+------------------------------------------+-------------------------+--------------------------+
10 rows in set (0.09 sec)
```

The provided SQL query displays the average and sum of benchmark points for each workflow type, with the result limited to the top 10.



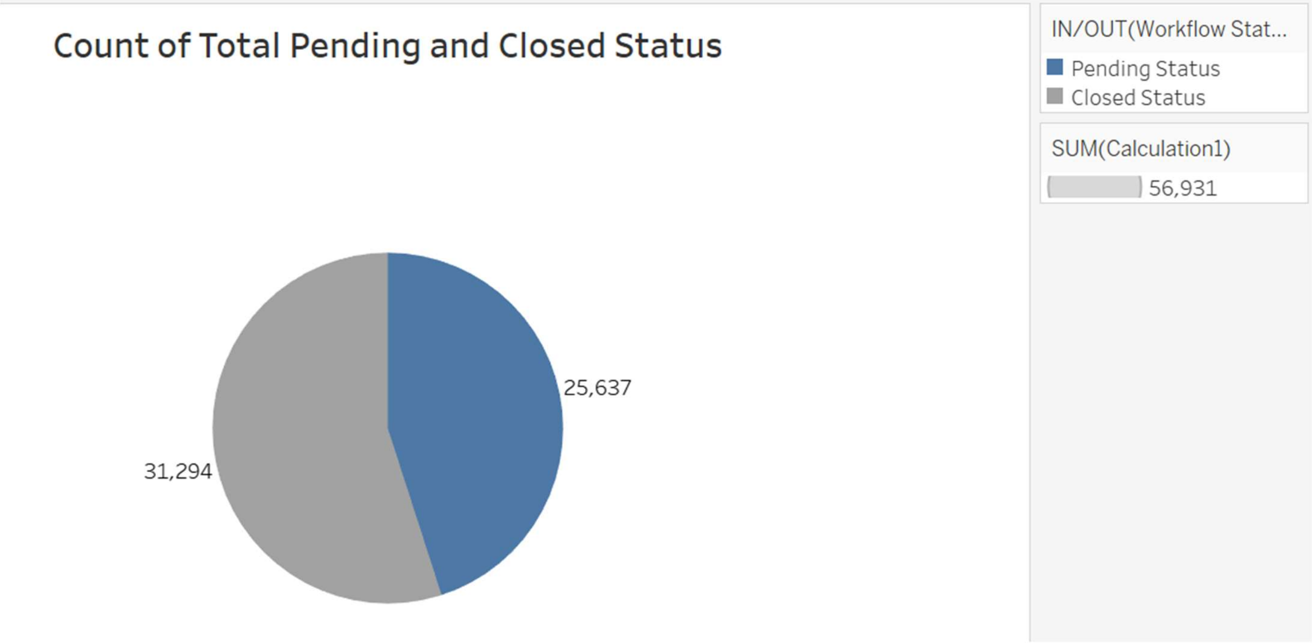
```
9. Different types of Workflow Status

mysql> select distinct(Workflow_status) as `Types of Workflow Status` from new_data;
+--------------------------+
| Types of Workflow Status |
+--------------------------+
| Pend QA                  |
| Closed                   |
| Pend Survey              |
| Pend Primary             |
| Pend Deletion            |
| Pend Correction          |
| Pend Secondary           |
+--------------------------+
7 rows in set (0.09 sec)
```

Above SQL query display the Different types of Workflow status

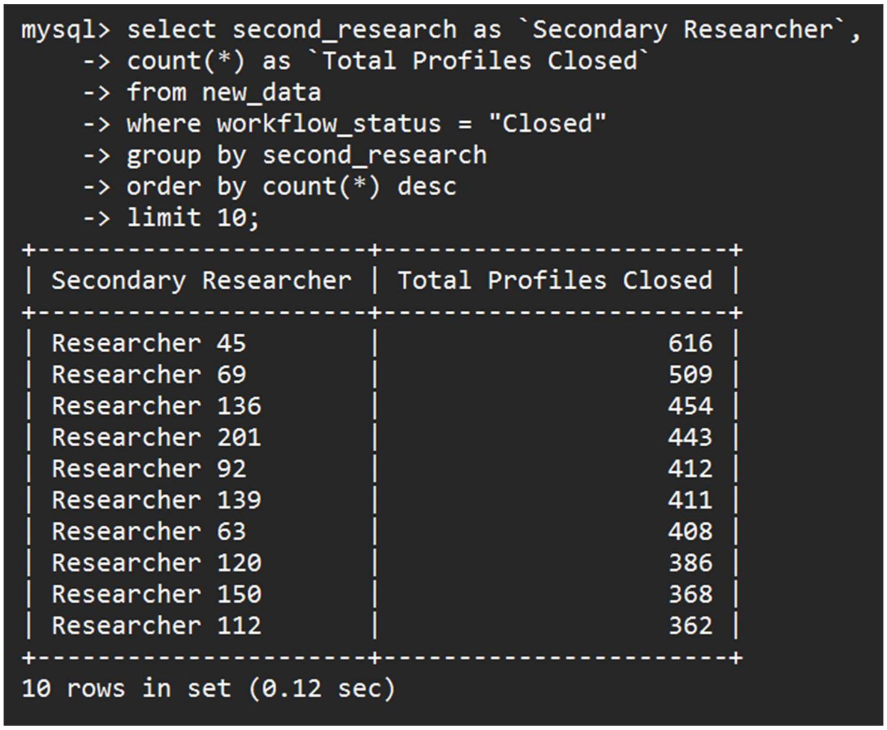


```
mysql> select count(*) As `Pending Profiles` from new_data
    -> where workflow_status in ("Pend QA","Pend Survey", "Pend Primary", "Pend Deletion", "Pend Correction", "Pend Secondary");
+------------------+
| Pending Profiles |
+------------------+
|            25637 |
+------------------+
1 row in set (0.05 sec)
```

Above SQL query shows the total pending profiles which includes workflow status as Pend QA, Pend Survey, Pend Primary, Pend Deletion, Pend Correction, Pend secondary

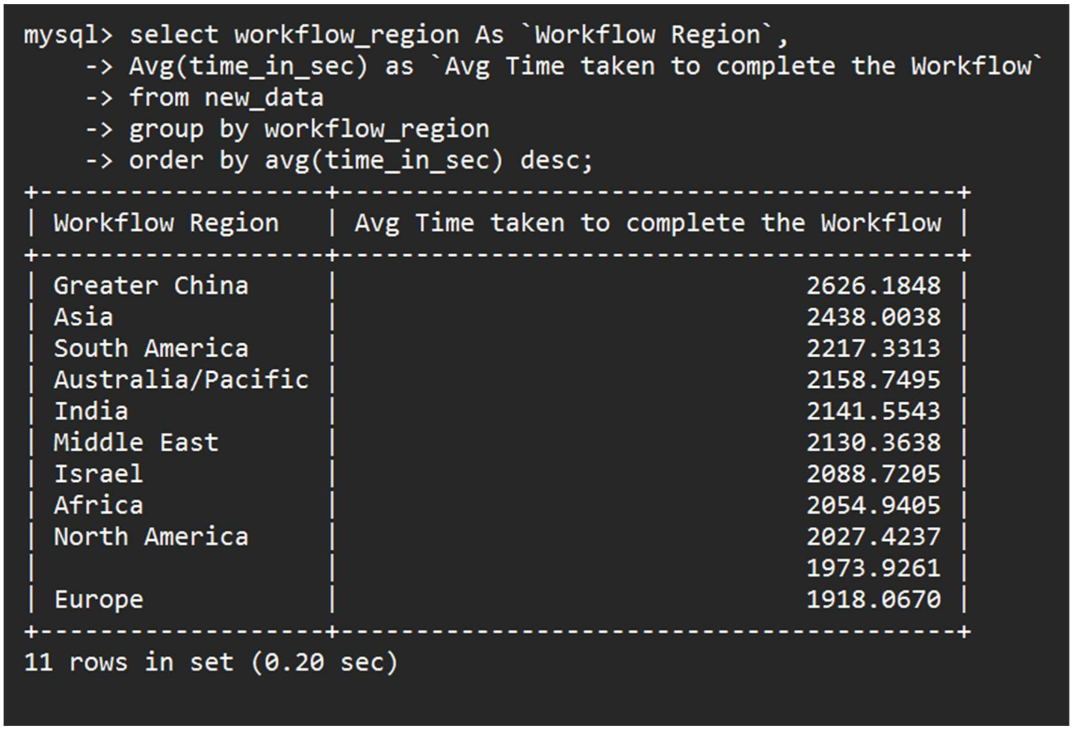


Above pie chart shows Sum of Pending/ Closed Workflow Status.



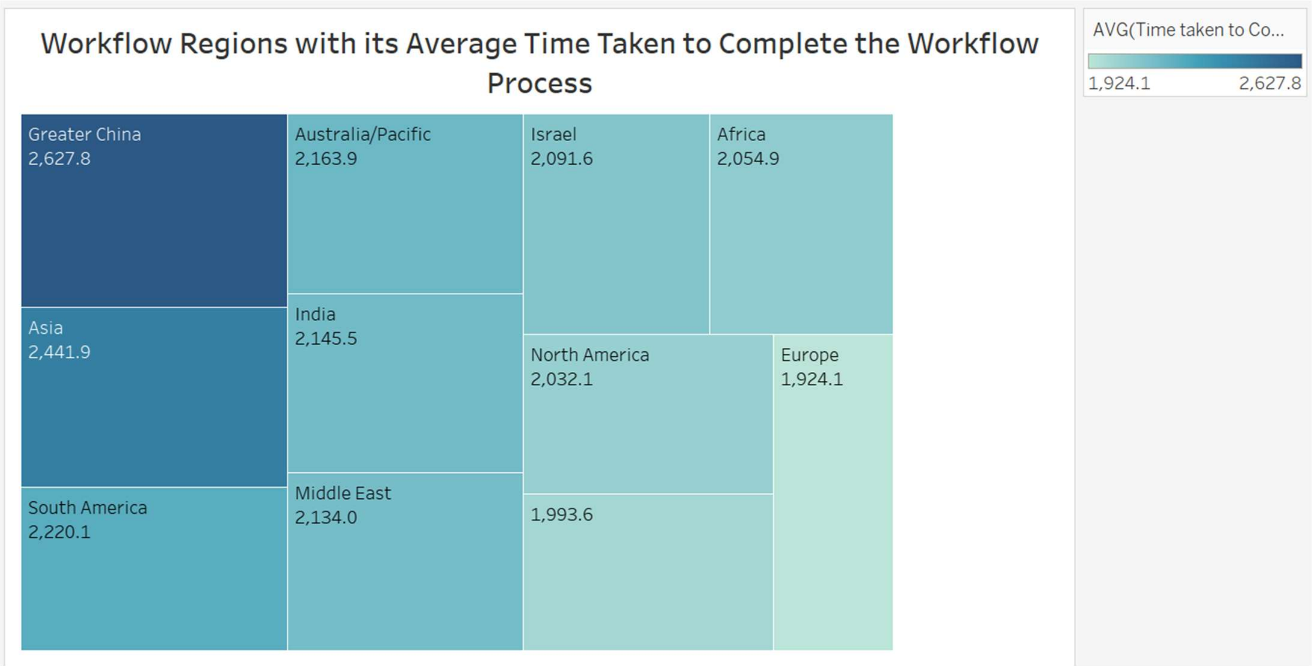
```
mysql> select second_research as `Secondary Researcher`,
    -> count(*) as `Total Profiles Closed`
    -> from new_data
    -> where workflow_status = "Closed"
    -> group by second_research
    -> order by count(*) desc
    -> limit 10;
+----------------------+-----------------------+
| Secondary Researcher | Total Profiles Closed |
+----------------------+-----------------------+
| Researcher 45        |                   616 |
| Researcher 69        |                   509 |
| Researcher 136       |                   454 |
| Researcher 201       |                   443 |
| Researcher 92        |                   412 |
| Researcher 139       |                   411 |
| Researcher 63        |                   408 |
| Researcher 120       |                   386 |
| Researcher 150       |                   368 |
| Researcher 112       |                   362 |
+----------------------+-----------------------+
10 rows in set (0.12 sec)
```

SQL query besides shows the Total Closed profiles by each Secondary researcher

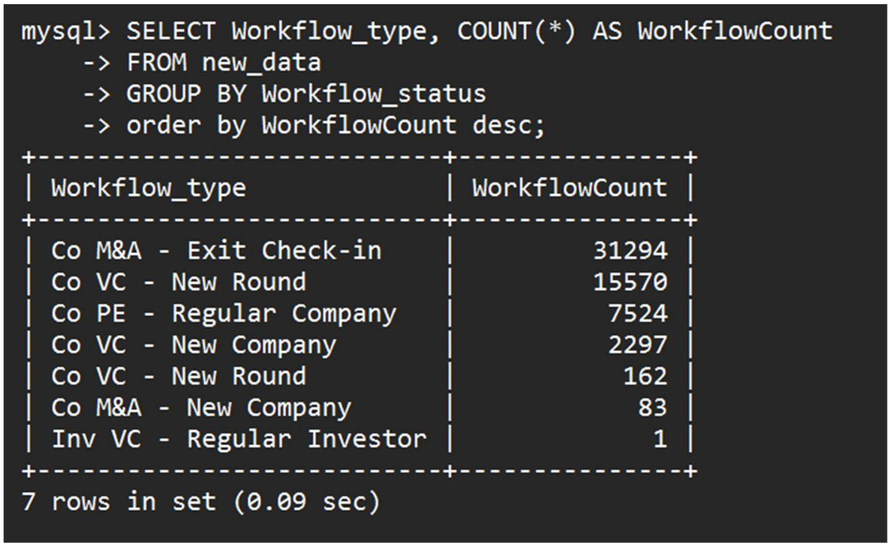


```
mysql> select workflow_region As `Workflow Region`,
    -> Avg(time_in_sec) as `Avg Time taken to complete the Workflow`
    -> from new_data
    -> group by workflow_region
    -> order by avg(time_in_sec) desc;
+-------------------+-----------------------------------------+
| Workflow Region   | Avg Time taken to complete the Workflow |
+-------------------+-----------------------------------------+
| Greater China     |                               2626.1848 |
| Asia              |                               2438.0038 |
| South America     |                               2217.3313 |
| Australia/Pacific |                               2158.7495 |
| India             |                               2141.5543 |
| Middle East       |                               2130.3638 |
| Israel            |                               2088.7205 |
| Africa            |                               2054.9405 |
| North America     |                               2027.4237 |
|                   |                               1973.9261 |
| Europe            |                               1918.0670 |
+-------------------+-----------------------------------------+
11 rows in set (0.20 sec)
```

Above SQL query shows Average time taken o complete the workflow process by each workflow region

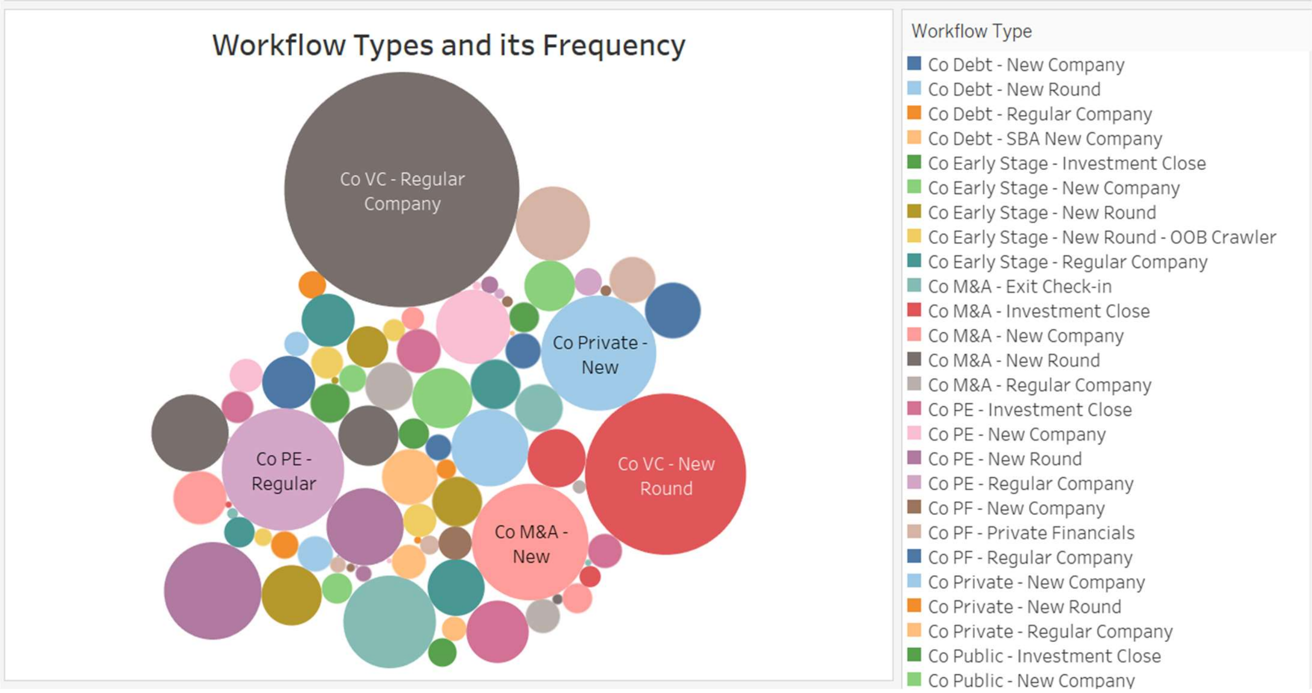


Workflow Region and average of Time taken to Complete (secs). Color shows average of Time taken to Complete (secs). Size shows average of Time taken to Complete (secs). The marks are labeled by Workflow Region and average of Time taken to Complete (secs).

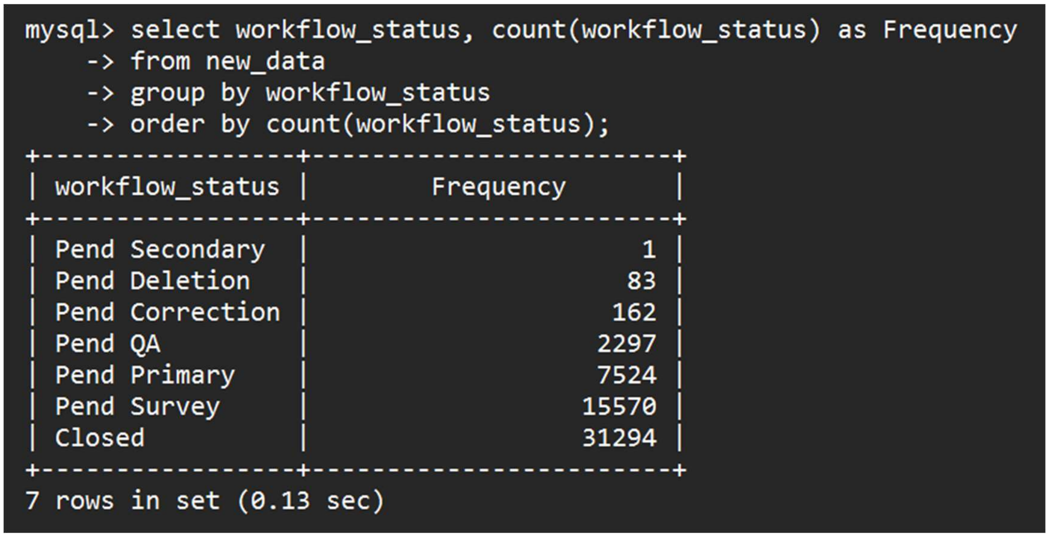


```
mysql> SELECT Workflow_type, COUNT(*) AS WorkflowCount
    -> FROM new_data
    -> GROUP BY Workflow_status
    -> order by WorkflowCount desc;
+---------------------------+---------------+
| Workflow_type             | WorkflowCount |
+---------------------------+---------------+
| Co M&A - Exit Check-in    |         31294 |
| Co VC - New Round         |         15570 |
| Co PE - Regular Company   |          7524 |
| Co VC - New Company       |          2297 |
| Co VC - New Round         |           162 |
| Co M&A - New Company      |            83 |
| Inv VC - Regular Investor |             1 |
+---------------------------+---------------+
7 rows in set (0.09 sec)
```

Above SQL query shows the Workflow Types and its frequency in dataset

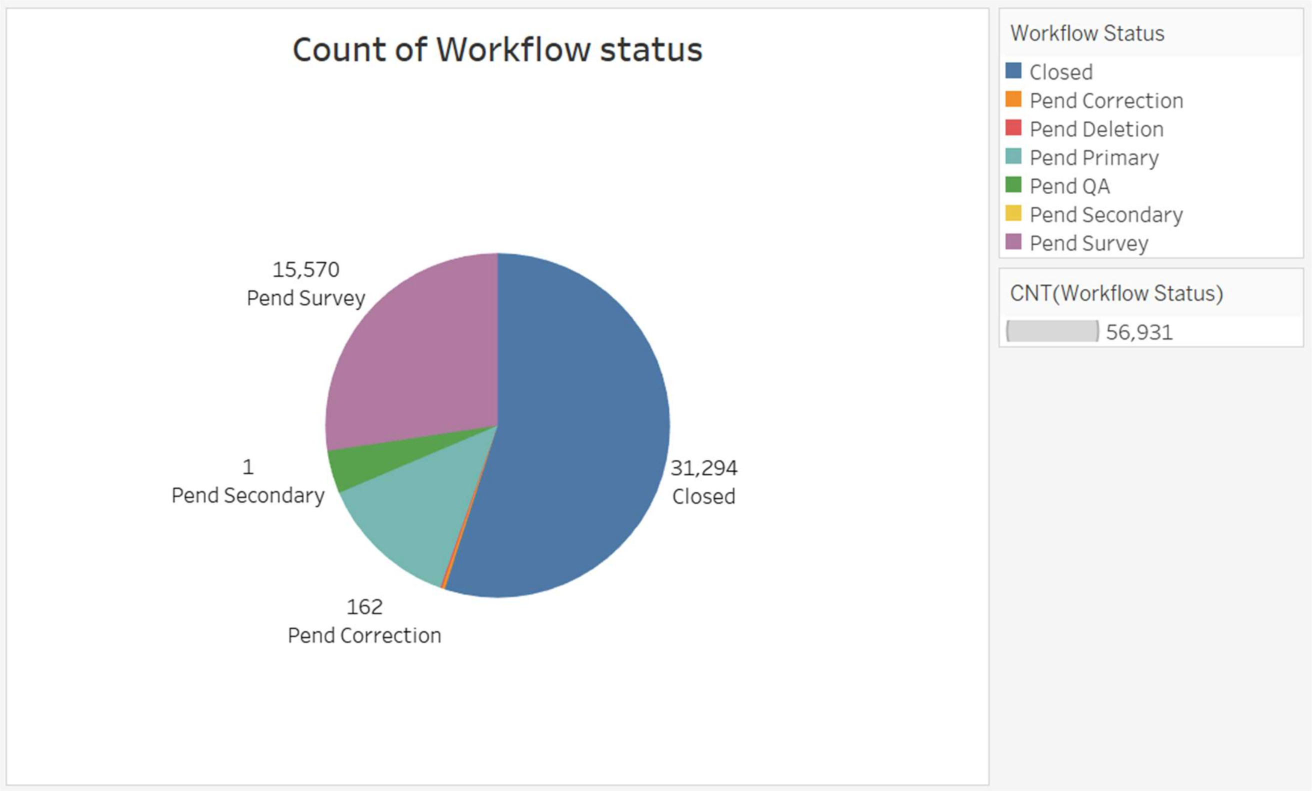


Workflow Type. Color shows details about Workflow Type. Size shows count of Workflow Type. The marks are labeled by Workflow Type.

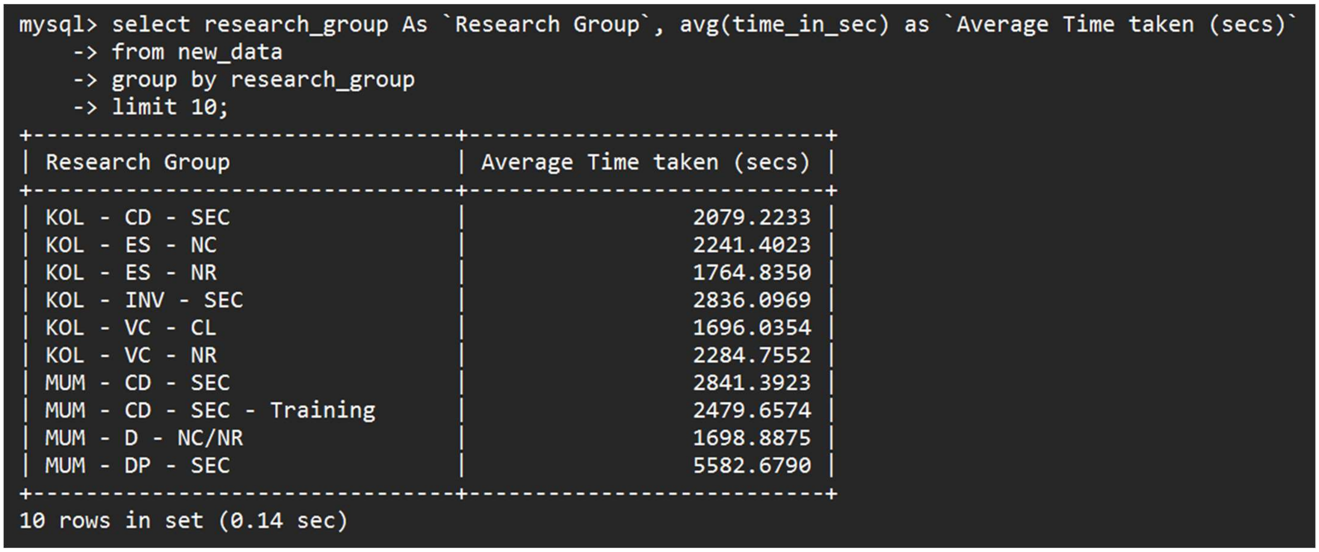


```
mysql> select workflow_status, count(workflow_status) as Frequency
    -> from new_data
    -> group by workflow_status
    -> order by count(workflow_status);
+----------------+----------------------+
| workflow_status |        Frequency       |
+----------------+----------------------+
| Pend Secondary |                    1 |
| Pend Deletion  |                   83 |
| Pend Correction |                  162 |
| Pend QA        |                 2297 |
| Pend Primary   |                 7524 |
| Pend Survey    |                15570 |
| Closed         |                31294 |
+----------------+----------------------+
7 rows in set (0.13 sec)
```

Above SQL Query shows Workflow Status and its frequency in dataset

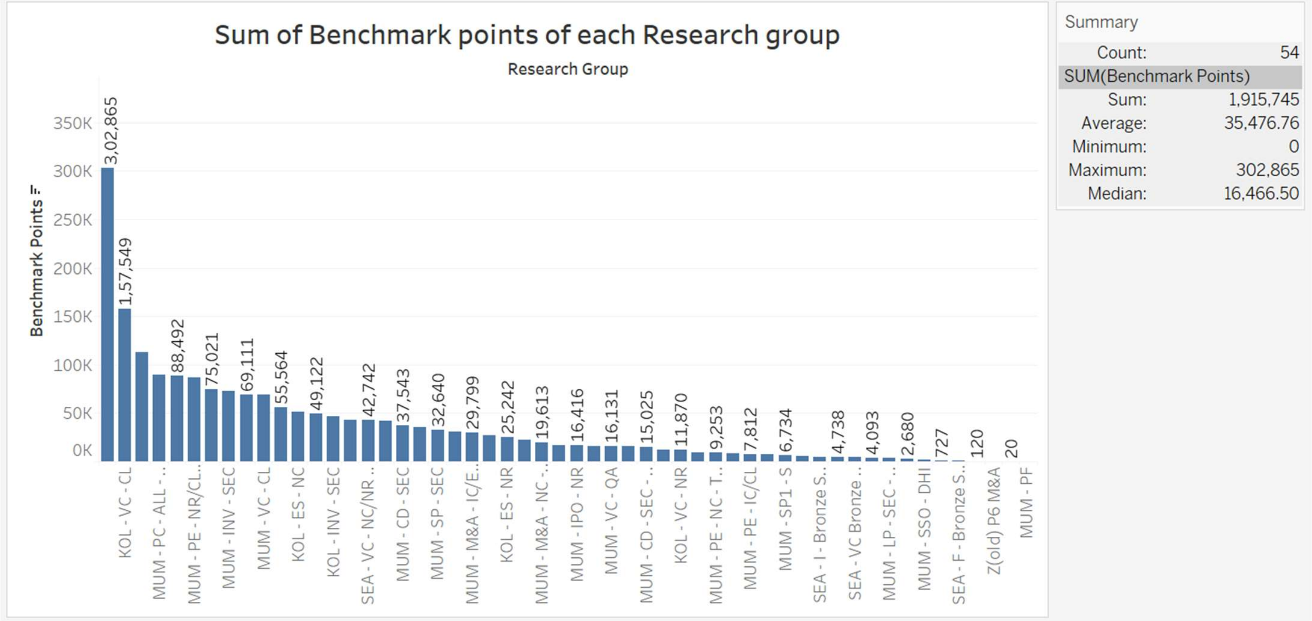


In the pie chart, the count of Workflow Status is represented by the size of each segment, and the color provides details about the respective Workflow Status. The marks on the chart are labeled with both the count and the specific Workflow Status. From this visualization, it is evident that the "Closed" workflow status has a higher frequency, followed by "Pend Survey".



```
mysql> select research_group As `Research Group`, avg(time_in_sec) as `Average Time taken (secs)`
    -> from new_data
    -> group by research_group
    -> limit 10;
+--------------------------------+---------------------------+
| Research Group                 | Average Time taken (secs) |
+--------------------------------+---------------------------+
| KOL - CD - SEC                 |                 2079.2233 |
| KOL - ES - NC                  |                 2241.4023 |
| KOL - ES - NR                  |                 1764.8350 |
| KOL - INV - SEC                |                 2836.0969 |
| KOL - VC - CL                  |                 1696.0354 |
| KOL - VC - NR                  |                 2284.7552 |
| MUM - CD - SEC                 |                 2841.3923 |
| MUM - CD - SEC - Training      |                 2479.6574 |
| MUM - D - NC/NR                |                 1698.8875 |
| MUM - DP - SEC                 |                 5582.6790 |
+--------------------------------+---------------------------+
10 rows in set (0.14 sec)
```

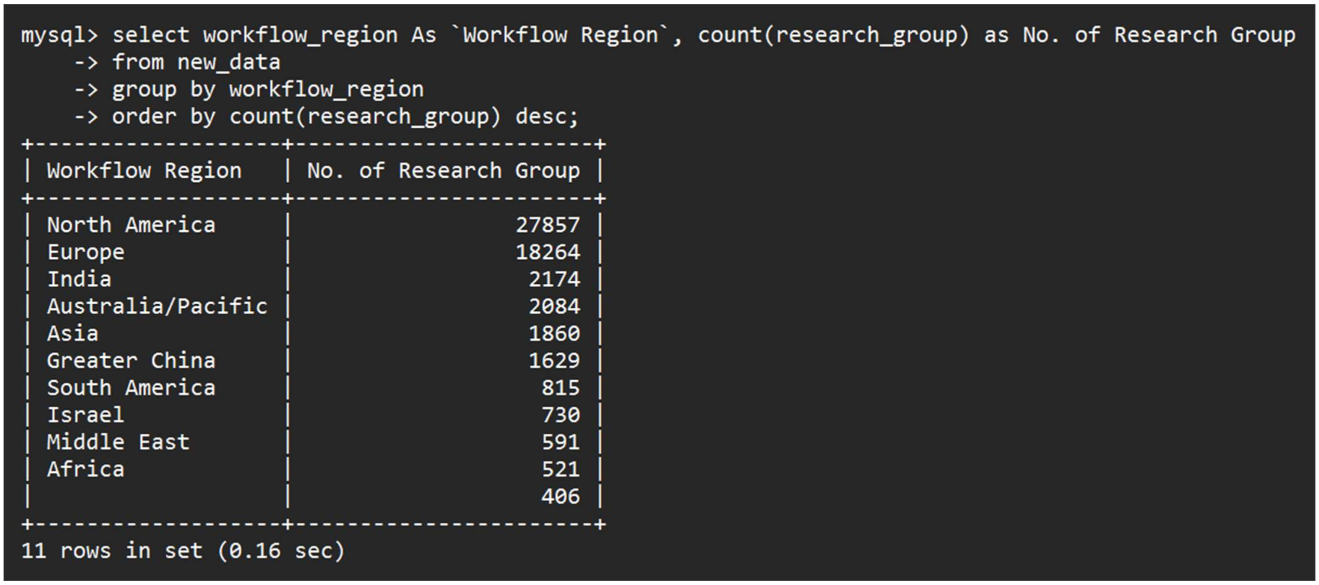
Above SQL query shows the Average of time taken by Each research group to complete the workflow process, result output is limited to 10.



Summary

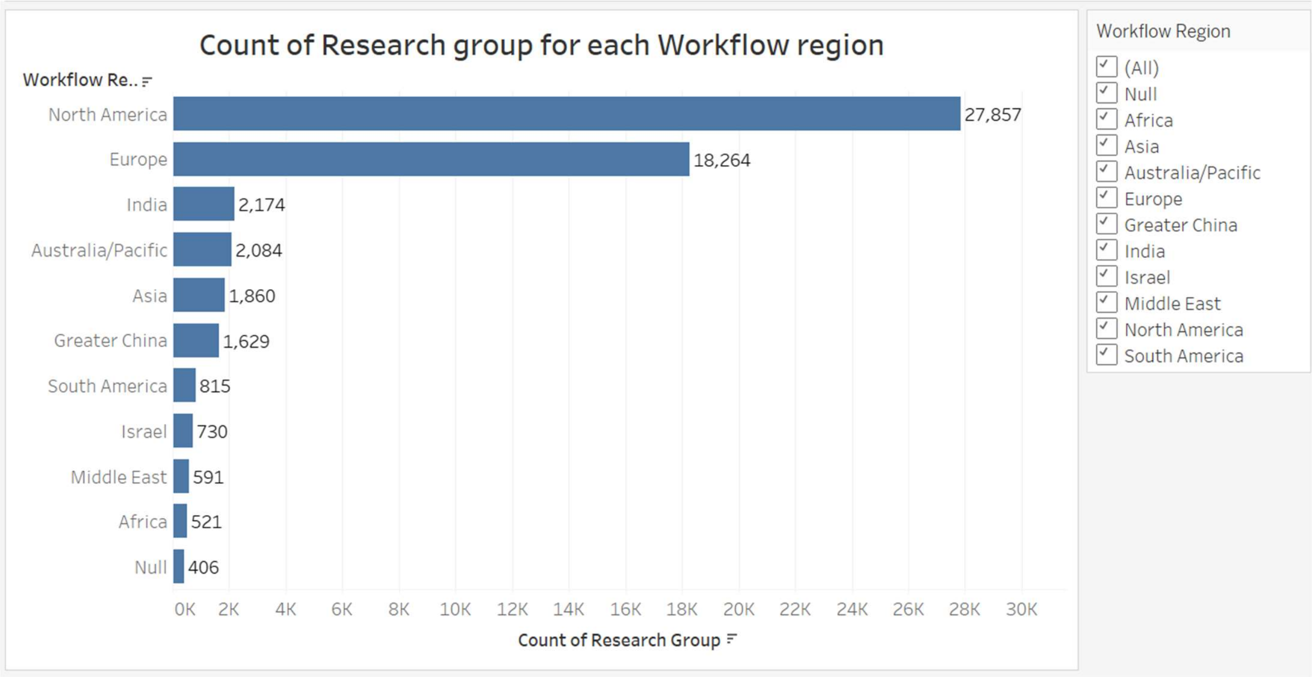
Count:	54
SUM(Benchmark Points)	
Sum:	1,915,745
Average:	35,476.76
Minimum:	0
Maximum:	302,865
Median:	16,466.50

The cumulative sum of Benchmark Points has been calculated for each Research Group, and the chart is annotated with the respective totals. The analysis indicates that the Research Group labeled as "MUM-VC-NR" has the highest sum of Benchmark Points compared to all other research groups.

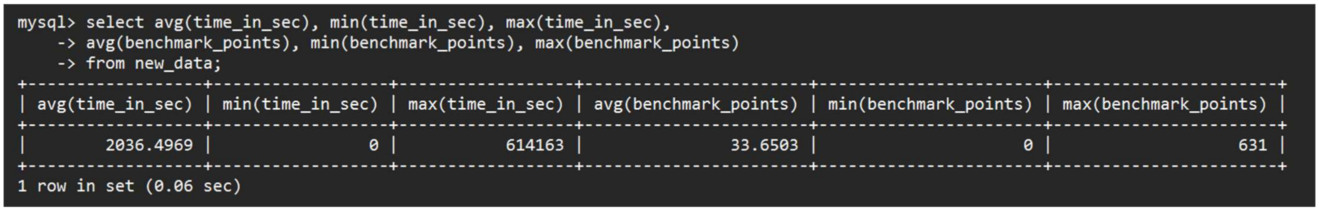


```
mysql> select workflow_region As `Workflow Region`, count(research_group) as No. of Research Group
    -> from new_data
    -> group by workflow_region
    -> order by count(research_group) desc;
+-------------------+-----------------------+
| Workflow Region   | No. of Research Group |
+-------------------+-----------------------+
| North America     |                 27857 |
| Europe            |                 18264 |
| India             |                  2174 |
| Australia/Pacific |                  2084 |
| Asia              |                  1860 |
| Greater China     |                  1629 |
| South America     |                   815 |
| Israel            |                   730 |
| Middle East       |                   591 |
| Africa            |                   521 |
|                   |                   406 |
+-------------------+-----------------------+
11 rows in set (0.16 sec)
```

Above SQL query No. of research groups there in each workflow region

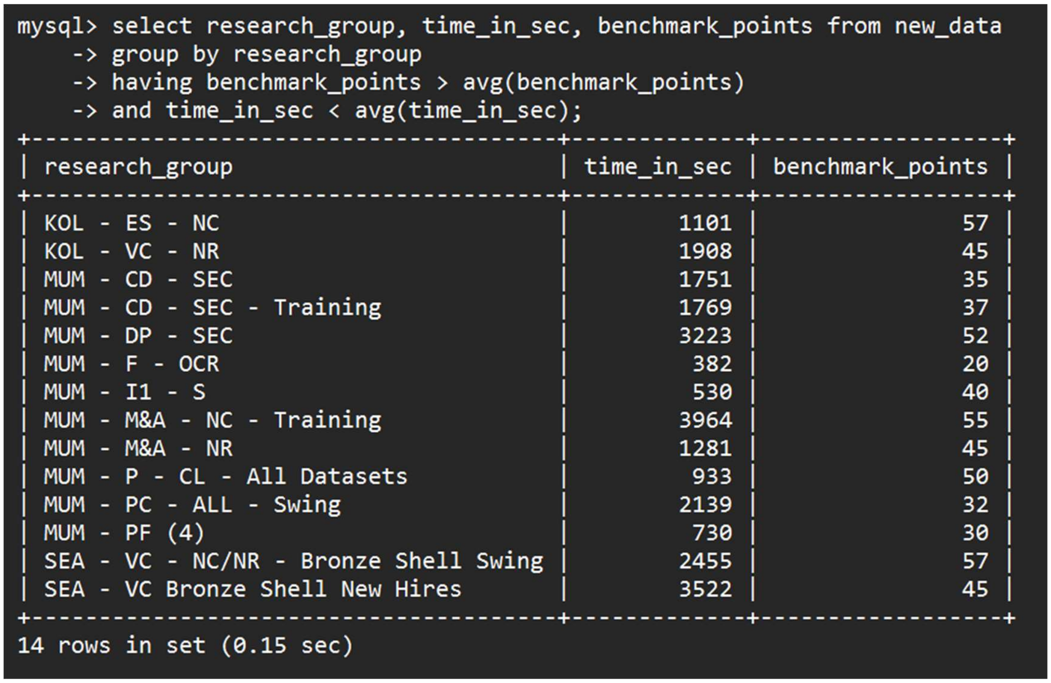


The pie chart displays the count of Research Groups for each Workflow Region, with marks labeled by the number of Research Groups. The view is filtered on Workflow Region, encompassing all 11 members. The visualization highlights that North America Region has a greater number of Research Groups compared to other regions.



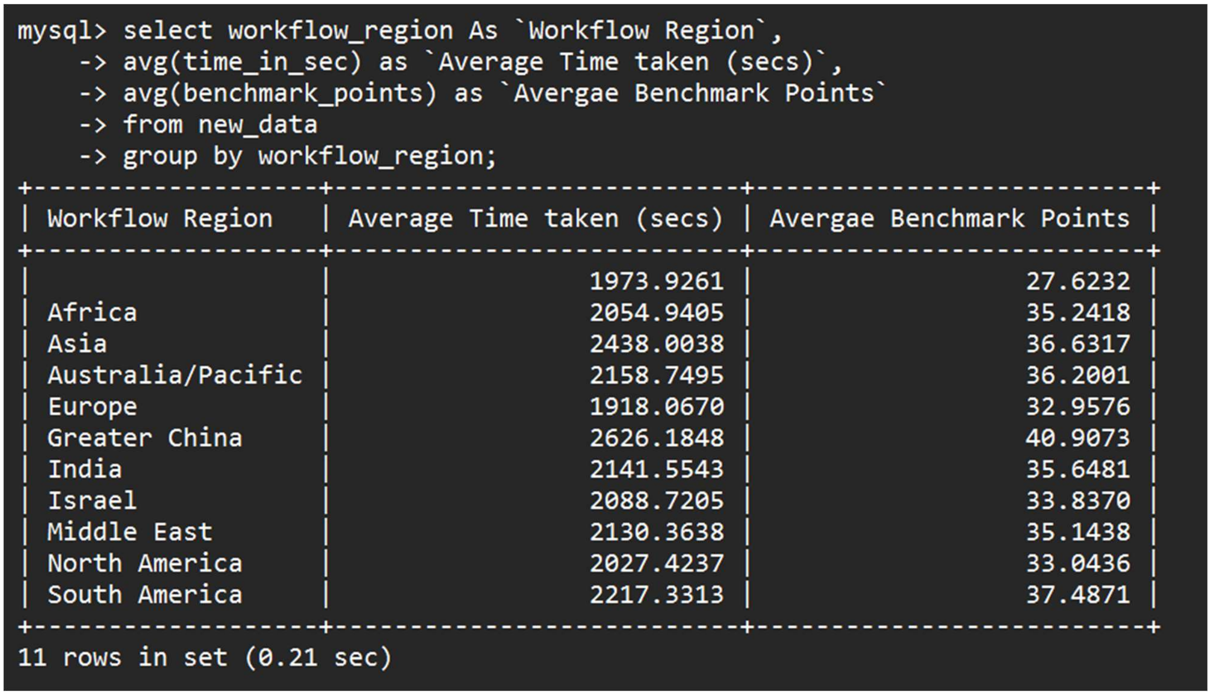
```
mysql> select avg(time_in_sec), min(time_in_sec), max(time_in_sec),
    -> avg(benchmark_points), min(benchmark_points), max(benchmark_points)
    -> from new_data;
+------------------+------------------+------------------+-----------------------+-----------------------+-----------------------+
| avg(time_in_sec) | min(time_in_sec) | max(time_in_sec) | avg(benchmark_points) | min(benchmark_points) | max(benchmark_points) |
+------------------+------------------+------------------+-----------------------+-----------------------+-----------------------+
|        2036.4969 |                0 |           614163 |               33.6503 |                     0 |                   631 |
+------------------+------------------+------------------+-----------------------+-----------------------+-----------------------+
1 row in set (0.06 sec)
```

Checking Min, Max and Average values of Numerical data



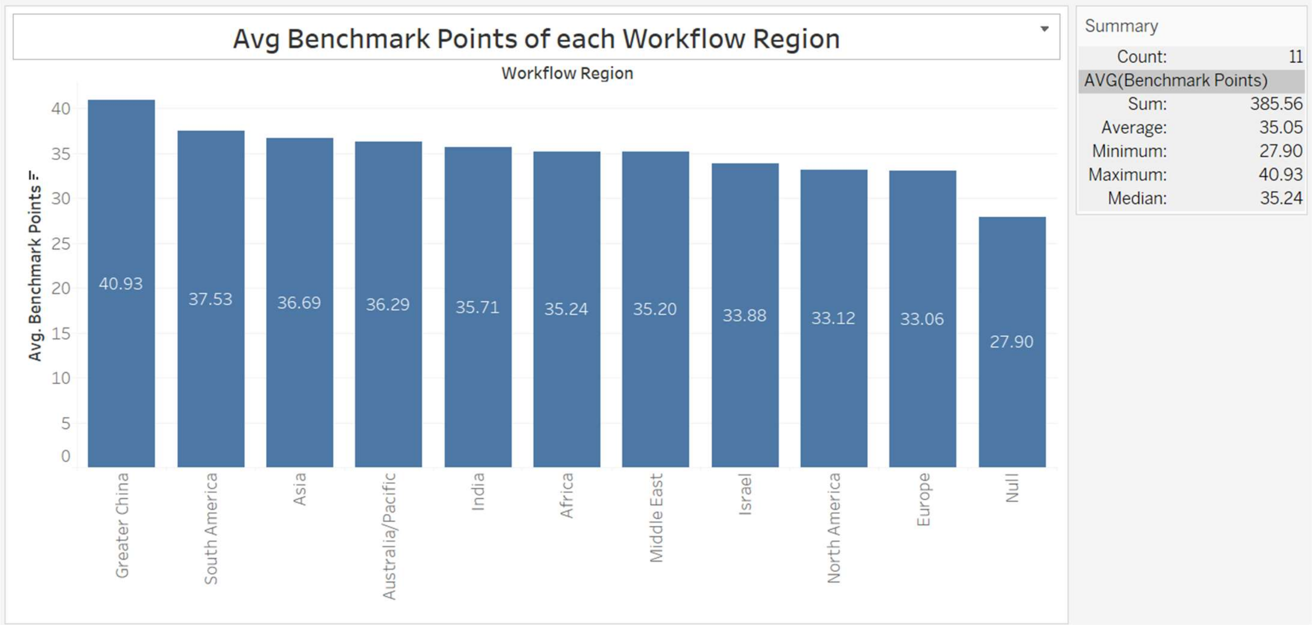
```
mysql> select research_group, time_in_sec, benchmark_points from new_data
    -> group by research_group
    -> having benchmark_points > avg(benchmark_points)
    -> and time_in_sec < avg(time_in_sec);
+-------------------------------------+------------+------------------+
| research_group                      | time_in_sec | benchmark_points |
+-------------------------------------+------------+------------------+
| KOL - ES - NC                       |        1101 |               57 |
| KOL - VC - NR                       |        1908 |               45 |
| MUM - CD - SEC                      |        1751 |               35 |
| MUM - CD - SEC - Training           |        1769 |               37 |
| MUM - DP - SEC                      |        3223 |               52 |
| MUM - F - OCR                       |         382 |               20 |
| MUM - I1 - S                        |         530 |               40 |
| MUM - M&A - NC - Training           |        3964 |               55 |
| MUM - M&A - NR                      |        1281 |               45 |
| MUM - P - CL - All Datasets         |         933 |               50 |
| MUM - PC - ALL - Swing              |        2139 |               32 |
| MUM - PF (4)                        |         730 |               30 |
| SEA - VC - NC/NR - Bronze Shell Swing |      2455 |               57 |
| SEA - VC Bronze Shell New Hires     |        3522 |               45 |
+-------------------------------------+------------+------------------+
14 rows in set (0.15 sec)
```

Above SQL query shows Efficiency of research groups which has less than average time taken and greater than average benchmark points



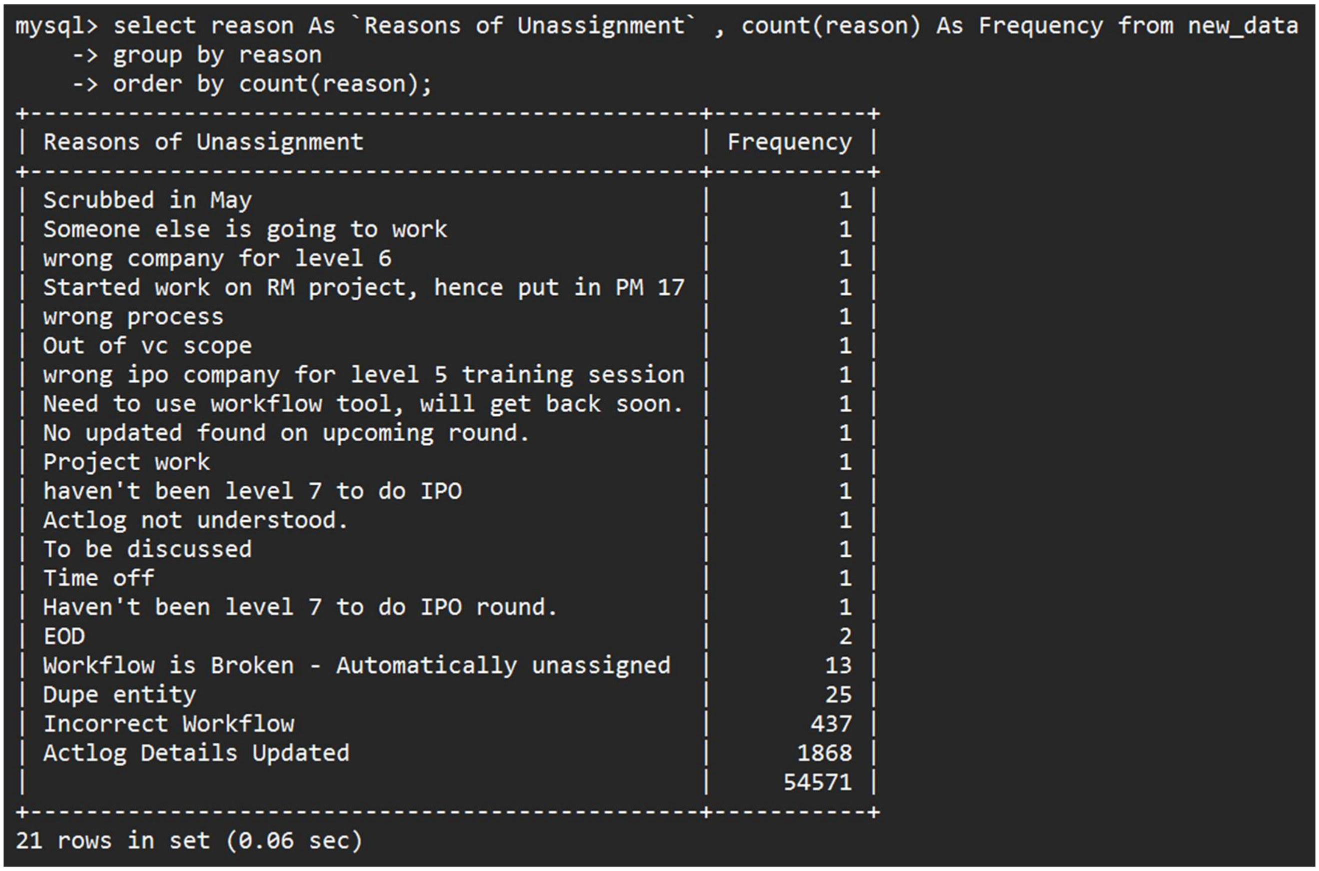
```
mysql> select workflow_region As `Workflow Region`,
    -> avg(time_in_sec) as `Average Time taken (secs)`,
    -> avg(benchmark_points) as `Avergae Benchmark Points`
    -> from new_data
    -> group by workflow_region;
+-------------------+---------------------------+--------------------------+
| Workflow Region   | Average Time taken (secs) | Avergae Benchmark Points |
+-------------------+---------------------------+--------------------------+
|                   |                 1973.9261 |                  27.6232 |
| Africa            |                 2054.9405 |                  35.2418 |
| Asia              |                 2438.0038 |                  36.6317 |
| Australia/Pacific |                 2158.7495 |                  36.2001 |
| Europe            |                 1918.0670 |                  32.9576 |
| Greater China     |                 2626.1848 |                  40.9073 |
| India             |                 2141.5543 |                  35.6481 |
| Israel            |                 2088.7205 |                  33.8370 |
| Middle East       |                 2130.3638 |                  35.1438 |
| North America     |                 2027.4237 |                  33.0436 |
| South America     |                 2217.3313 |                  37.4871 |
+-------------------+---------------------------+--------------------------+
11 rows in set (0.21 sec)
```

Average time taken and average benchmark points by each workflow region



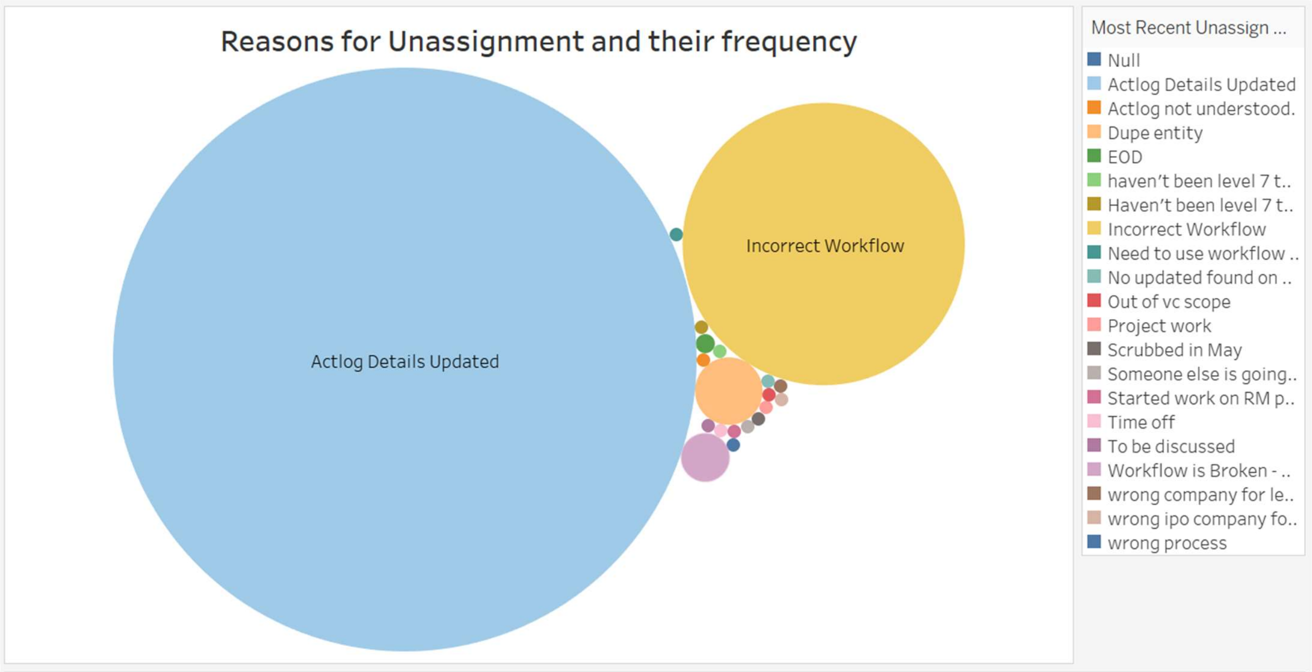
Summary	
Count:	11
AVG(Benchmark Points)	
Sum:	385.56
Average:	35.05
Minimum:	27.90
Maximum:	40.93
Median:	35.24

Average of Benchmark Points for each Workflow Region. The marks are labeled by average of Benchmark Points.

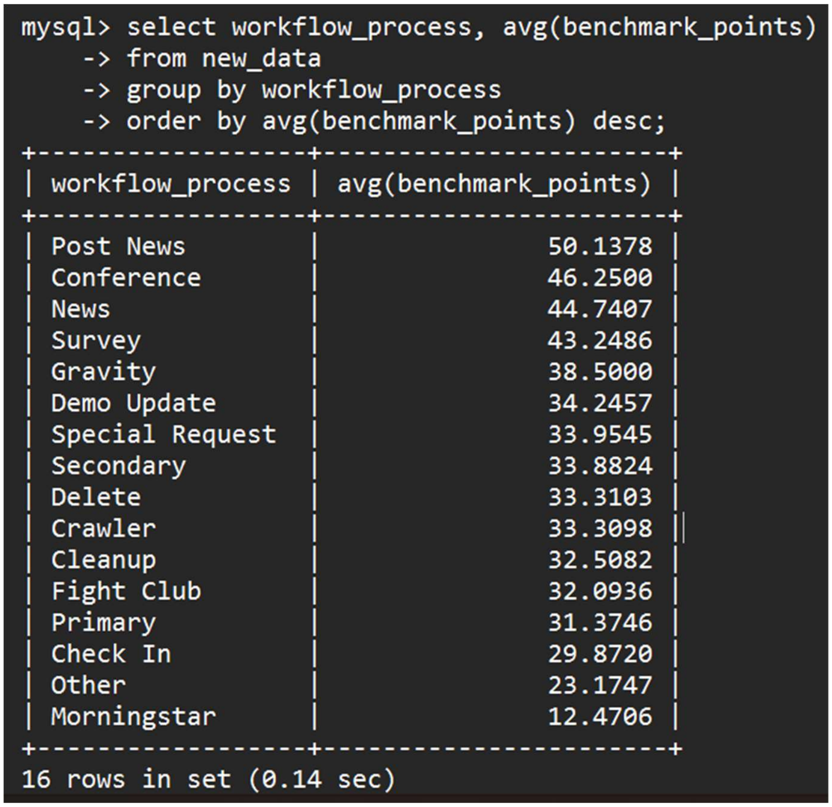


```
mysql> select reason As `Reasons of Unassignment` , count(reason) As Frequency from new_data
    -> group by reason
    -> order by count(reason);
+-------------------------------------------------+-----------+
| Reasons of Unassignment                         | Frequency |
+-------------------------------------------------+-----------+
| Scrubbed in May                                 |         1 |
| Someone else is going to work                   |         1 |
| wrong company for level 6                       |         1 |
| Started work on RM project, hence put in PM 17  |         1 |
| wrong process                                   |         1 |
| Out of vc scope                                 |         1 |
| wrong ipo company for level 5 training session  |         1 |
| Need to use workflow tool, will get back soon.  |         1 |
| No updated found on upcoming round.             |         1 |
| Project work                                    |         1 |
| haven't been level 7 to do IPO                  |         1 |
| Actlog not understood.                          |         1 |
| To be discussed                                 |         1 |
| Time off                                        |         1 |
| Haven't been level 7 to do IPO round.           |         1 |
| EOD                                             |         2 |
| Workflow is Broken - Automatically unassigned   |        13 |
| Dupe entity                                     |        25 |
| Incorrect Workflow                              |       437 |
| Actlog Details Updated                          |      1868 |
|                                                 |     54571 |
+-------------------------------------------------+-----------+
21 rows in set (0.06 sec)
```

Reasons of Unassignment and its frequency

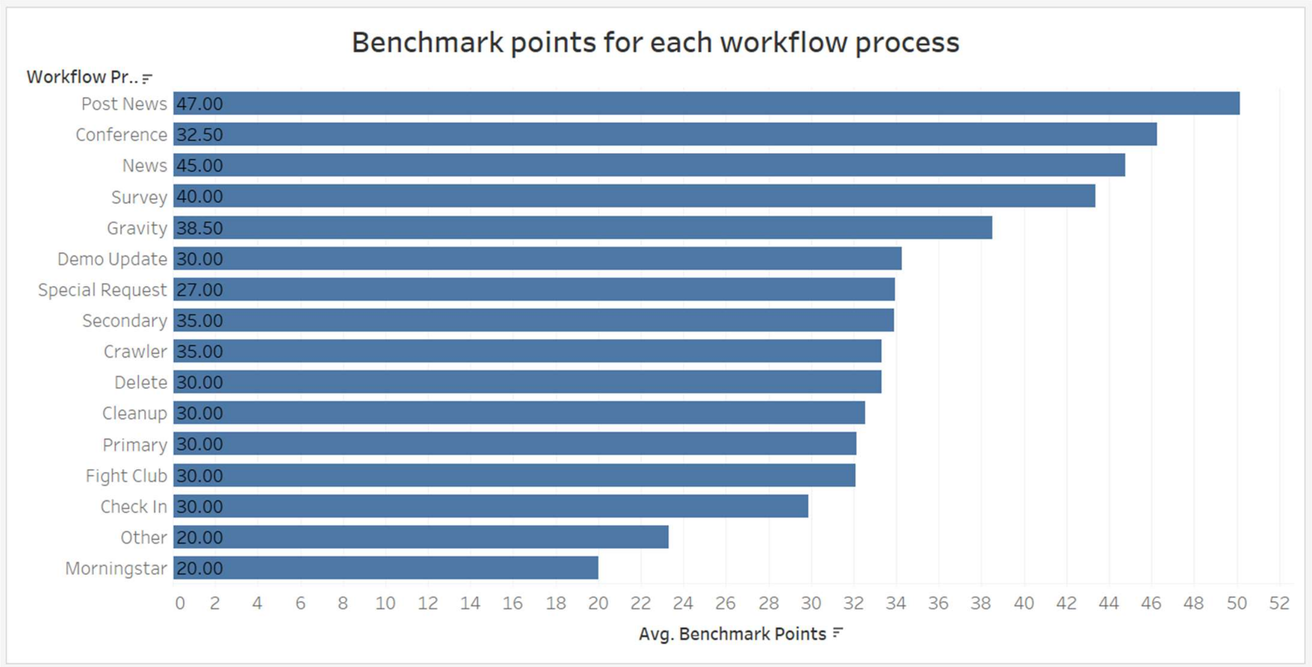


The bubble chart illustrates the distribution of Most Recent Unassign Reason, with bubble size indicating the count of each reason. The labels on the marks reveal that "Actlog Details Updated" is the most frequently encountered reason, closely followed by "Incorrect Workflow".

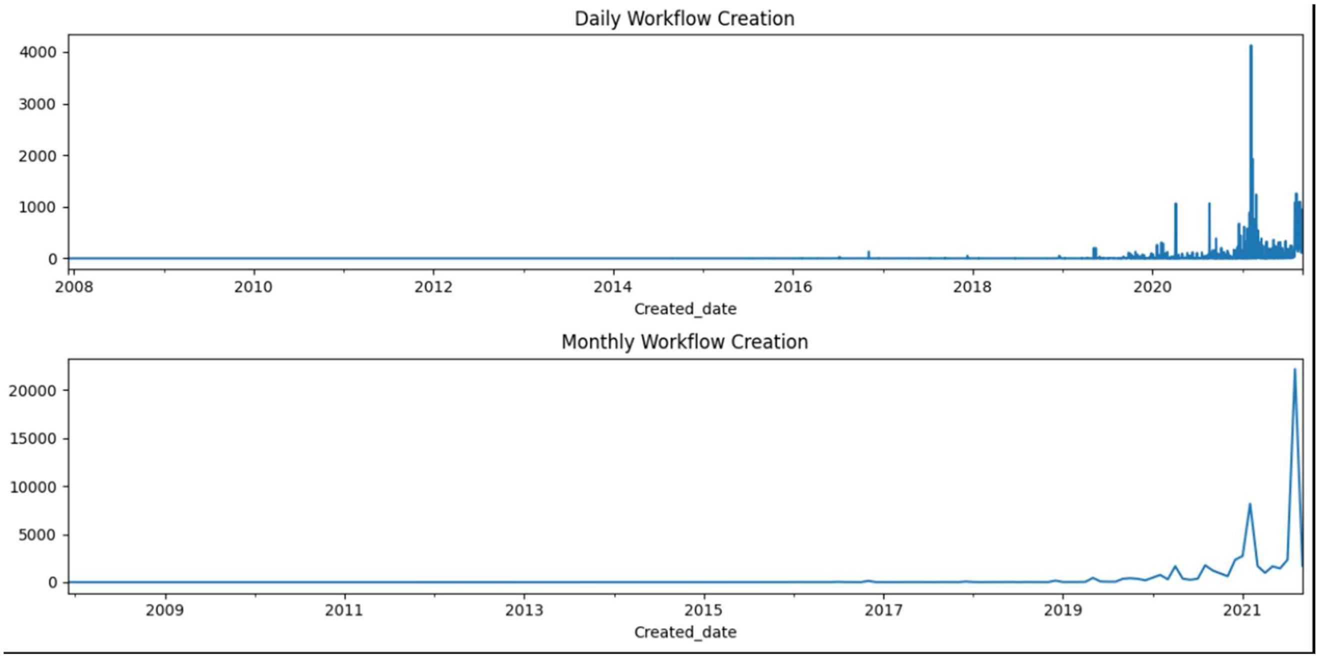


```
mysql> select workflow_process, avg(benchmark_points)
    -> from new_data
    -> group by workflow_process
    -> order by avg(benchmark_points) desc;
+------------------+-----------------------+
| workflow_process | avg(benchmark_points) |
+------------------+-----------------------+
| Post News        |               50.1378 |
| Conference       |               46.2500 |
| News             |               44.7407 |
| Survey           |               43.2486 |
| Gravity          |               38.5000 |
| Demo Update      |               34.2457 |
| Special Request  |               33.9545 |
| Secondary        |               33.8824 |
| Delete           |               33.3103 |
| Crawler          |               33.3098 |
| Cleanup          |               32.5082 |
| Fight Club       |               32.0936 |
| Primary          |               31.3746 |
| Check In         |               29.8720 |
| Other            |               23.1747 |
| Morningstar      |               12.4706 |
+------------------+-----------------------+
16 rows in set (0.14 sec)
```

Average Benchmark Points of each Workflow Process



The chart illustrates the average Benchmark Points for each Workflow Process. It is evident from the chart that the "Post News" Workflow process has a higher average of Benchmark Points compared to other workflow processes.



Daily and Monthly Workflow

Tools used: SQL, Python (Pandas, Seaborn, Matplotlib, Ydata_profiling), Tableau, MS Word

Thank You
Manoj Gaikwad
Manoj20497@gmail.com
+91 8286930944