Datum Group

Group Members

Students ID	Students Name	Section
442002988	Monerah Almobarak	91S
442005104	Sarah Aljuhani	91S
442000786	Sarah Altaweel	91S
442003374	Nada Alotaibi	91S

Table of Contents

Table of Tables

Table of Figures

1.0 Introduction:

Datum College is a computer college specializing in data and artificial intelligence (AI). The dataset we have is focused on the faculty members at Datum College. It provides information about each faculty member, including their ID numbers, ranks, years of experience, and corresponding salaries. This dataset is essential for gaining insights into the composition of the faculty and understanding how they are remunerated. Analyzing this data in-depth can provide valuable information for assessing the organizational structure of the faculty, identifying potential patterns or trends, and supporting decision-making processes related to faculty recruitment and compensation. By exploring this dataset comprehensively, we aim to uncover meaningful insights that contribute to the continuous improvement and effective functioning of Datum College's faculty.

2.0 Team Members Roles and Contribution:

The table provides an overview of the team members' roles and contributions to the project.

Team Member	Roles and Contribution
Monerah Almobarak	Introduction, Conclusions, Question 6
Sarah Aljuhani	Question1, Question3
Sarah Altaweel	Question2, Question7
Nada Alotaibi	Questions 4, Questions 5, Logo making

Table 1.Team Members Roles and Contributions.

3.0 Q 1. Provide a brief description of the dataset (population, observations, variables' types).

- **str() function**: This function provides a compact way to display the structure of an R object, including its type, dimensions, and contents.

```
'data.frame':   403 obs. of  4 variables:
 $ ID        : int  1 2 3 4 5 6 7 8 9 10 ...
 $ Rank      : chr  "Prof" "Prof" "AsstProf" "Prof" ...
 $ Experience: int  18 16 3 39 41 6 23 45 20 18 ...
 $ Salary    : int  139750 173200 79750 115000 141500 97000 175000 147765 119250 129000 ...
```

Figure 1.Output of the str() function.

The output from this function is a data frame with '403' observations and '4 'variables.

- **summary() function**: This function provides a summary of the main characteristics of the variables in a dataset, such as minimum and maximum values, mean, median, and quartiles.

```
       ID            Rank           Experience
Min.   :  1.0   Length:403         Min.   :  0.00
1st Qu.:100.5   Class :character   1st Qu.:  7.00
Median :201.0   Mode  :character   Median : 16.50
Mean   :200.4                      Mean   : 17.95
3rd Qu.:300.5                      3rd Qu.: 27.00
Max.   :400.0                      Max.   :150.00
                                   NA's   :3

    Salary
Min.   : 57800
1st Qu.: 91025
Median :107175
Mean   :113478
3rd Qu.:133975
Max.   :231545
NA's   :1
```

Figure 2.Output of the summary() function.

When we run summary() on a data frame, it will provide the following information:

- Mean - The average of all numeric columns.
- Standard deviation - The measure of dispersion of a numeric column from its mean.
- Minimum value - The minimum value in each numeric column.
- Maximum value - The maximum value in each numeric column.

- **head() function**: This function displays the first few rows of a dataset

	ID <int>	Rank <chr>	Experience <int>	Salary <int>
1	1	Prof	18	139750
2	2	Prof	16	173200
3	3	AsstProf	3	79750
4	4	Prof	39	115000
5	5	Prof	41	141500
6	6	AssocProf	6	97000

6 rows

Figure 3. Output of the head() function.

- **dim() function:** This function provides the dimensions of a dataset, which can be useful for determining how many rows and columns the data has.

```
[1] 403   4
```

Figure 4. Output of the dim() function.

This is a dimension of a dataset with '403' rows and '4' columns.

- **names() function**: This function displays the names of the variables in a dataset, which can be useful for identifying which columns contain which information.

```
[1] "ID"         "Rank"       "Experience" "Salary"
```

Figure 5. Output of the names() function.

4.0 Q 2 Consider the quality factors and provide a quality report on the raw data.

We considered six quality factors:

1- Uniqueness

Uniqueness is a crucial aspect when assessing the quality of a dataset, as it directly impacts data accuracy, reliability, and overall integrity. Ensuring uniqueness means eliminating duplicate entries, where each data instance represents a distinct and singular observation.

The significance of uniqueness cannot be overstated, as it influences various aspects of data analysis and management. Duplicate entries undermine the distinctiveness and individuality of data instances, introducing inconsistencies and potential errors. By evaluating the presence of duplicates, we specifically address the quality dimension of uniqueness and its importance in maintaining data accuracy.[1]

Detecting and removing duplicate entries is essential to mitigate the risk of incorrect conclusions, misleading trends, and unreliable insights. It allows for a more accurate and reliable analysis, enhancing the overall quality of the dataset. By eliminating duplicate rows and ensuring unique identification through the ID column, we promote data integrity and enhance the reliability of our findings.

In our case, the initial dataset consisted of 403 rows. After data cleansing successfully removed both duplicate rows and duplicates based on the ID column, we streamlined the dataset to 400 rows. This indicates that our efforts effectively eliminated duplicate entries, resulting in a dataset that upholds the desired level of uniqueness.

To summarize, the importance of uniqueness lies in its role in preserving data integrity, ensuring accurate analysis, and facilitating reliable decision-making. By addressing uniqueness and actively removing duplicate entries, we establish a solid foundation for robust and trustworthy data exploration and interpretation.

2- Completeness

Completeness is a critical aspect of data quality, and it refers to the presence of all expected and necessary values in a dataset. In our code, we specifically focused on assessing the completeness dimension by checking for the presence of missing values.

Identifying missing values in a dataset is crucial as it indicates potential gaps or incomplete information, which can significantly impact the reliability and validity of data analysis. In our analysis, we discovered that the dataset contains 4 missing values, emphasizing the need for further data cleaning.[2]

The presence of missing values raises concerns about the representativeness of the dataset and can lead to biased or inaccurate analysis results. Missing values may arise due to various reasons, such as data collection errors, data entry mistakes, or respondents choosing not to

provide certain information. Regardless of the cause, it is essential to address missing values to ensure a complete and reliable dataset.

Handling missing values requires thoughtful consideration and appropriate techniques, such as imputation or exclusion methods. By addressing missing values, we improve the completeness of the dataset, which enhances the accuracy and reliability of subsequent analysis and decision-making.

Complete data enables us to draw more robust conclusions, make meaningful comparisons, and derive accurate insights. They provide a more accurate representation of the underlying population and contribute to a more comprehensive understanding of the phenomena under study.

In summary, completeness plays a vital role in data quality. Detecting and addressing missing values is crucial for ensuring the reliability of analysis outcomes. By acknowledging the presence of 4 missing values in the dataset, we highlight the importance of data cleaning to enhance completeness and maximize the usefulness of the data for further analysis.[3]

3- Validity

Our code also focuses on the quality dimension of data validity, specifically addressing outliers, as the third quality factor. By utilizing a box plot, we visually identify potential outliers in the dataset. This allows us to assess the presence of data points that significantly deviate from the norm and may raise concerns about the validity of the data.

To identify outliers, we apply Tukey's fences method, which defines the range within which values are considered typical. This enables us to detect and handle outliers that exhibit substantial deviations from the expected data patterns. It becomes apparent that there are outliers present in both the Salary and Experience attributes.

In the case of the Salary attribute, the identified outliers provide valuable insights into potential wage gaps or discriminatory practices within the dataset. These extreme values contribute to a comprehensive understanding of the distribution and potential inequities within the salary data. Removing these outliers may not be the most appropriate course of action, as they hold crucial information that can help uncover disparities and contribute to further analysis and investigation.

On the other hand, outliers in the Experience attribute require careful consideration due to their potential impact on the analysis. While outliers can offer insights into unusual cases or data entry errors, extreme outliers can introduce significant distortions and compromise the reliability of our results. For example, an instance with an experience of 150 years is highly unlikely and is likely an erroneous data point. Addressing such extreme outliers is necessary to ensure the integrity and robustness of our analysis.

By removing these extreme outliers from the Experience attribute, we can mitigate their adverse effects and obtain more accurate insights into the relationship between experience and other factors within the dataset. This helps to ensure that our analysis is based on valid and reliable data, free from significant distortions caused by extreme outliers.

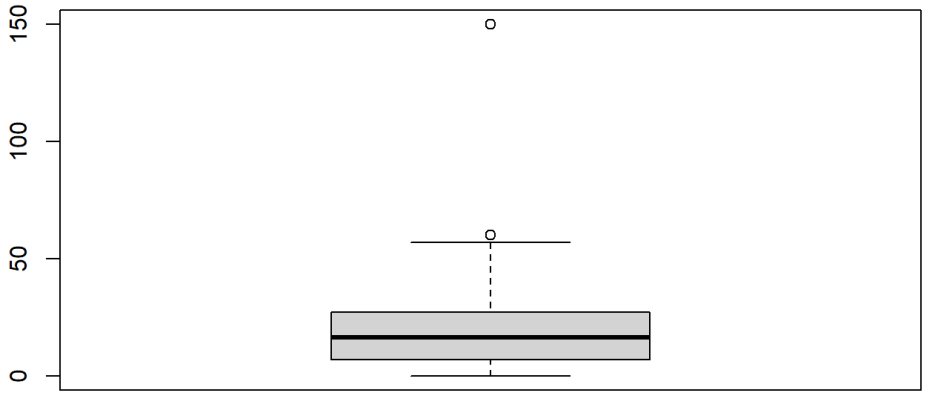

Figure 6. Output of the boxplot() function before removing outliers in faculty experience.

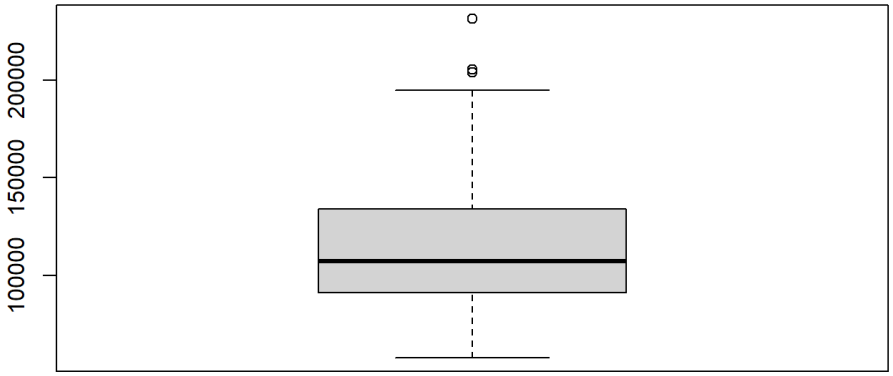

Figure 7. Output of the boxplot() function before removing outliers in faculty salary .

4- Consistency

Consistency plays a crucial role in maintaining the correctness and uniformity of attribute values within a dataset. In the context of our analysis, we focused on the quality dimension of consistency, specifically targeting potentially misspelled ranks. By identifying and addressing misspelled ranks, we aimed to enhance data consistency and correctness.

Misspelled ranks can introduce inconsistencies and errors in the dataset, affecting the reliability and accuracy of data analysis and reporting. In our examination, we discovered two instances of misspelled ranks: "AssstProf" and "AssocProff". These misspelled entries may indicate data entry errors or inconsistencies in the recording of ranks.

Ensuring consistency in attribute values is vital for various reasons. First, it improves the reliability and integrity of the dataset, enabling more accurate analysis and decision-making. By identifying and correcting misspelled ranks, we eliminate potential sources of error and enhance the overall quality of the dataset.

Second, maintaining consistent attribute values supports data integration and interoperability. When datasets from different sources are combined or compared, having consistent attribute values ensures compatibility and facilitates effective data integration. By addressing

misspelled ranks, we promote consistency, making it easier to merge and align datasets for comprehensive analysis.

Furthermore, consistent attribute values facilitate efficient data filtering, searching, and querying. When attribute values are spelled consistently, it simplifies the process of retrieving specific data based on criteria or conducting advanced searches. By rectifying misspelled ranks, we improve data handling capabilities and enable more accurate and efficient data retrieval.

In summary, our focus on identifying and addressing misspelled ranks underscores the importance of consistency in data quality. By rectifying inconsistencies and errors, we enhance the dataset's reliability, integrity, and usability for robust analysis and decision-making.[4]

5- Relevancy

The relevancy of data is a crucial quality factor that ensures the information captured is meaningful, applicable, and aligns with the objectives of the analysis. In the context of our dataset, we can confidently assert that the data possesses relevancy, making it valuable for our analysis and decision-making processes.

The dataset comprises attributes such as ID, Salary, Experience, and Rank, which are essential elements for capturing pertinent information about faculty members. These attributes provide key insights into unique identifiers, salary levels, years of service, and the positions held within the academic hierarchy. The data's relevance stems from its ability to contribute meaningful information for analysing salary distributions, identifying experience trends, and understanding the composition of different ranks among faculty members.

Moreover, the data's relevancy is reinforced by its source, which is a reliable provider. The credibility of the data source ensures the accuracy and integrity of the information, further enhancing its relevancy. By relying on a trusted source, we can have confidence in the data's authenticity, reducing concerns about potential biases or inaccuracies.

The relevancy of the data is instrumental in conducting various analyses and making informed decisions. It allows us to explore salary distributions among faculty members, identify patterns or disparities based on experience, and gain insights into the distribution of different ranks within the academic institution. The relevance of the data enables us to draw meaningful conclusions, inform policies or strategies, and address pertinent questions related to faculty compensation, career progression, and organizational structure.

In summary, the data's relevancy is a fundamental quality dimension that ensures its applicability and significance for our analysis. The attributes captured in the dataset, sourced from a reliable provider, offer valuable insights into faculty-related aspects, allowing us to make informed decisions and gain a comprehensive understanding of various factors influencing faculty members' roles and compensation.[6]

6- Timeliness

The timeliness of data is a significant quality factor that should be considered when analyzing the dataset. In the context of this dataset, it is important to recognize that the data may lack timeliness and may not capture the most recent or up-to-date information.

Timeliness is a crucial aspect of data quality as it ensures the relevance and accuracy of the dataset. By acknowledging the timeliness dimension, we understand that the dataset provides a snapshot of the attributes (ID, Salary, Experience, and Rank) at a specific point in time. It represents a historical perspective rather than reflecting the current state of the faculty members' attributes.

While timeliness is desirable for some analyses, it may not be a primary concern for others. Depending on the research question or objective, historical data can still provide valuable insights and contribute to a comprehensive understanding of trends, patterns, or historical comparisons.

However, it is important to be aware that the dataset's timeliness may have implications for certain analyses or decision-making processes that require real-time or recent data. It is essential to consider the potential impact of data timeliness on the validity and applicability of the results and conclusions drawn from the analysis.

In summary, the timeliness of the data is a critical quality dimension to consider in this dataset. Acknowledging the dataset's historical nature and its limitations in capturing the most recent information is essential for the proper interpretation and contextualization of the analysis results. By recognizing the timeliness dimension, we can appropriately assess the data's applicability and make informed decisions based on the dataset's temporal context.

5.0 Q 3 Apply required operations for data cleansing.

1-Removing duplicates: Use the unique() function :remove duplicate rows in the dataset.

	ID <int>	Rank <chr>	Experience <int>	Salary <int>
1	1	Prof	18	139750
2	2	Prof	16	173200
3	3	AsstProf	3	79750
4	4	Prof	39	115000
5	5	Prof	41	141500
6	6	AssocProf	6	97000
7	7	Prof	23	175000
8	8	Prof	45	147765
9	9	Prof	20	119250
10	10	Prof	18	129000

1-10 of 401 rows Previous 1 2 3 4 5 6 ... 41 Next

Figure 8.Output of the updated dataset.

This is the output after removing duplicate rows which contain 400 rows.

2-Handling missing values:

1. we use na.omit() to create a new data frame cleaned_data with all rows containing NA (missing) values removed from the original Data data frame.
2. It then checks if the number of rows in cleaned_data is less than the original Data, using nrow(). If so, that means rows were removed, and a message is printed.
3. Finally, it assigns the cleaned_data data frame back to Data, overwriting the original data with the cleaned version.

```
[1] "Rows with missing values were removed."
```

Figure 9.Output after remove missing values.

3-Handling outliers:

we use the boxplot() function :to identify outliers in the Experience, and then decide how to handle them.

This Plot before removing outliers from Experience. We remove outliers her because the experience impossible to be more than 150 years.

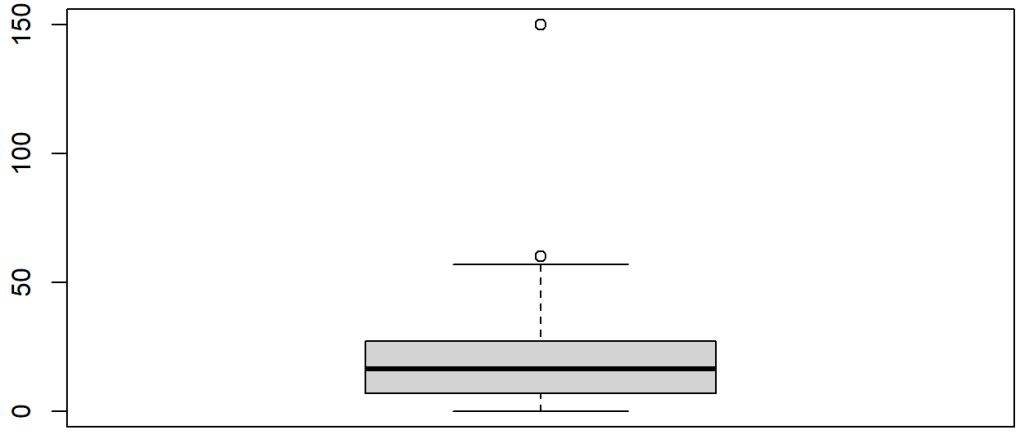

Figure 10. Output of the boxplot() function before removing outliers in faculty experience.

This plot after removing outliers from Experience:

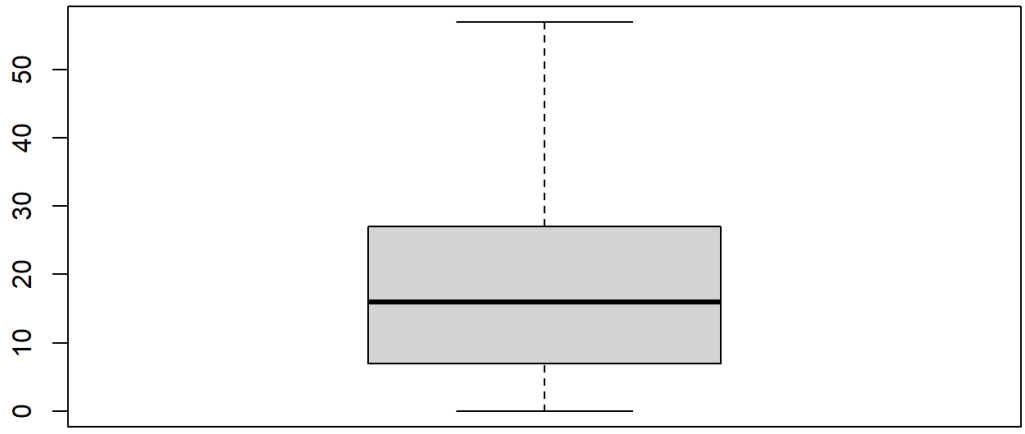

Figure 11. Output of the boxplot() function after removing outliers in faculty experience.

4-Correct misspell

This code addresses the data quality dimensions of consistency and correctness by:

1. Identifying misspelled ranks in the dataset using string distance calculations. It compares each rank to a list of correct spellings using the Jaro-Winkler distance method.
2. Correcting the misspelled ranks by replacing them with the closest correct spelling based on minimum string distance.
3. Storing the original misspelled rank and the corrected rank in a data frame to log the corrections made.
4. Updating the attribute in the dataset with the corrected ranks.
5. Printing a summary of the corrections to verify which misspelled ranks were found and corrected.

This approach has the following benefits for data quality:

- It ensures ranks are spelled consistently according to the correct spellings list.
- It fixes spelling errors and incorrect values, promoting data correctness.
- It utilizes an algorithmic, similarity-based approach to identifying and correcting misspellings automatically.
- It logs the corrections made for auditability and traceability.

Old_Rank <chr>	Corrected_Rank <chr>
AssstProf	AsstProf
AssocProff	AssocProf

2 rows

Figure 12. Output after correcting misspelling.

After running the code, it is observed that the dataset contains misspelled ranks, which have now been corrected. By utilizing similarity-based correction, the misspelled ranks were replaced with the most appropriate and accurate values. This correction enhances the consistency and correctness of attribute values.

6.0 Q4 Provide appropriate plots for each attribute or variable.

1- Plotting Rank attribute by using bar chart plot:

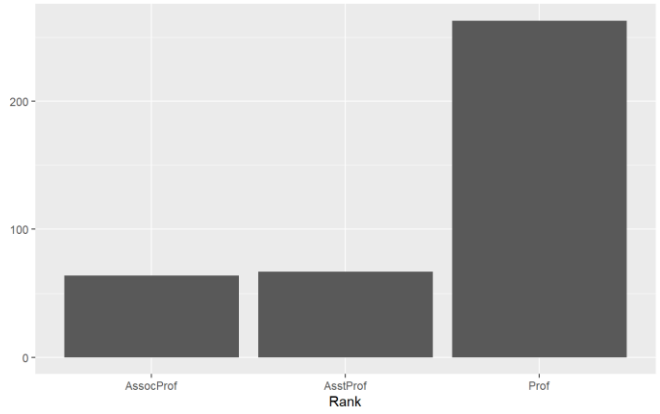

Figure 13. Faculty Ranking by bar chart.

It became clear to us that most of the faculty college members are professors.

2- Plotting Experience attribute by using histogram plot:

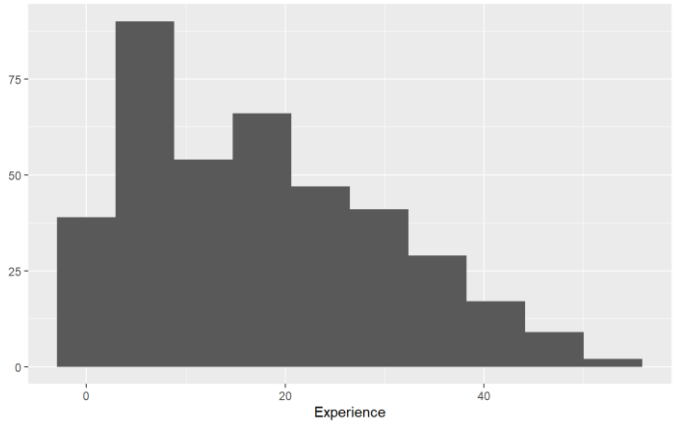

Figure 14. Faculty experience by histogram.

The experience has appeared that have right-skewed distributions.

3- Plotting Salary attribute by using histogram plot:

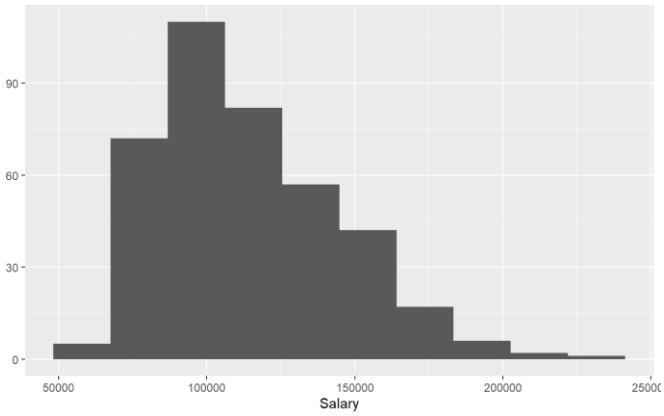

Figure 15. Faculty salary by histogram.

The salary has appeared that have right-skewed distributions.

7.0 Q5. Provide appropriate plots that visualize relations or associations between each pair of variables.

1- Plotting associations between Rank and Salary attributes by using a Hexagonal heatmap plot:

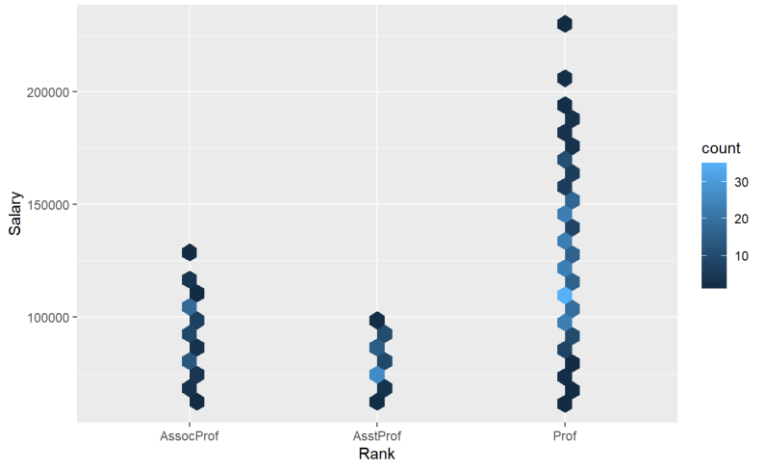

Figure 16. Associations between faculty ranking and salaries - hexagonal heatmap

It turns out that the Professor has the highest salary and Associate Professor has less than the Professor amount of salary and the Assistant professor has the lowest salary.

2- Plotting associations between Rank and Experience attributes by using a Hexagonal heatmap plot:

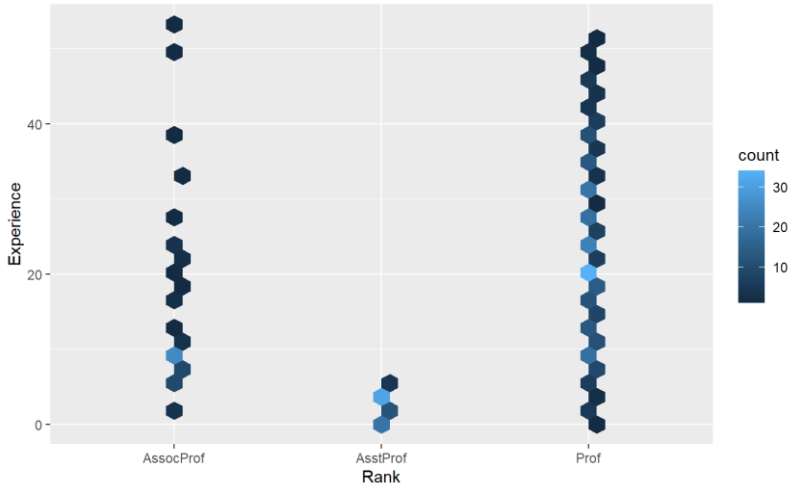

Figure 17. Associations between faculty ranking and experience - hexagonal heatmap.

As it turns out, the assistant professor has the shortest experience time, followed by the associate professor and the professor with the longest experience.

3- Plotting associations between Experience and Salary attributes by using a Hexagonal heatmap plot:

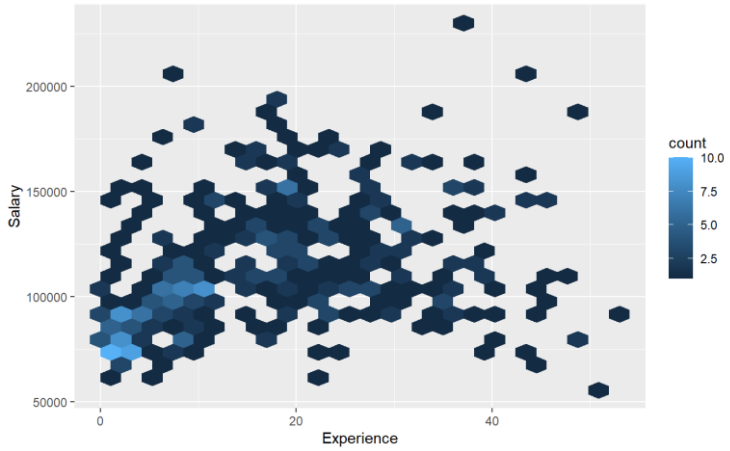

Figure 18.Associations between faculty salaries and experience - hexagonal heatmap.

4- The Hexagonal heatmap graphic indicates that the majority of the faculty college members are paid almost between $50,000 and $150,000.

5- Plotting associations between the whole three attributes Experience, Salary, and Rank attributes by using a scatter plot:

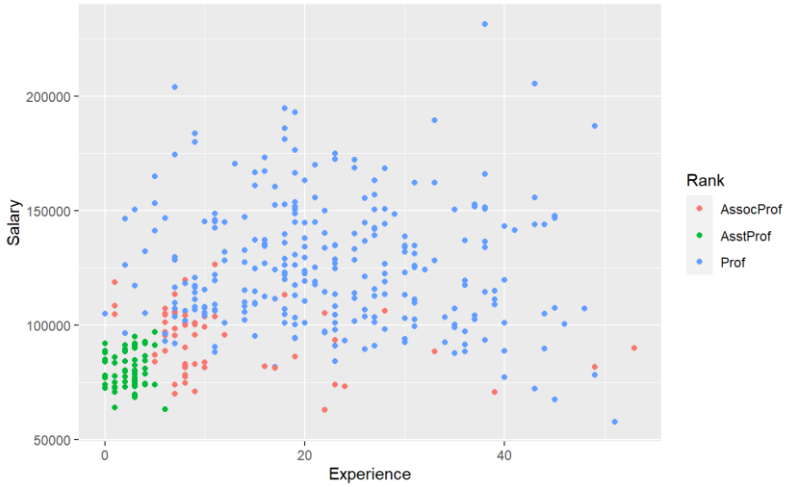

Figure 19. Associations between faculty ranking, salaries,and experience - scatter plot.

If we examine the incomes of the three-rank category, we discover that:

1- The assistant professor position is centred on a certain number of experiences, with the salary being broadly rounded from those with no experience to those who have at least five years of experience and a salary of up to 100,000.

2- The associate Professor has a variety of experiences, but most of them are based on having fewer than 20 years of experience, and they earn around 103,000 each.

3- The professors with the longest tenures and most experience earn approximately between 100,000 and 153,000.

8.0 Q6 Are there any discriminations or wage gaps that are not justified?

a. Does "faculty-rank" affect "faculty-salary"? Justify

yes, faculty rank seems to influence faculty salary. professors earn higher salaries compared to associate professors and assistant professors. This observation is supported by the box plot, which illustrates the differences in salary levels among various faculty ranks.

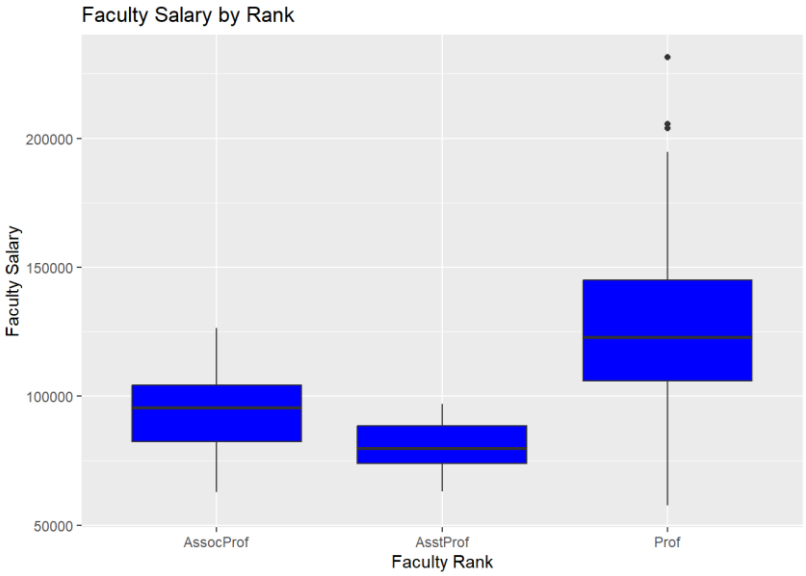

Figure 20. Faculty Salary by Rank using boxplot.

b. Does "Faculty-Experience" impact "faculty salary"? Justify

yes, faculty experience seems to have an impact on faculty salary. As experience increases, faculty members generally tend to earn higher salaries. This relationship can be effectively visualized using a scatter plot and a linear regression line, which helps to demonstrate the positive correlation between experience and salary.

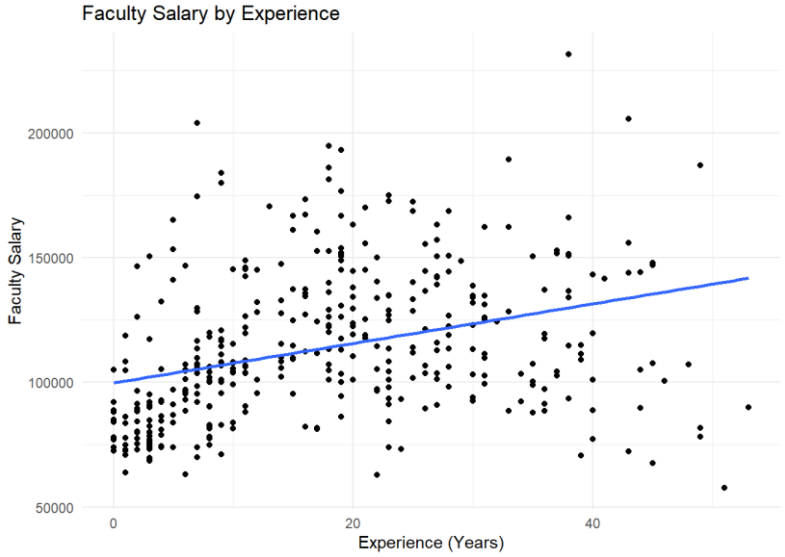

Figure 21. Faculty Salary by Experience using scatter plot.

c. Is the difference between associate professors' salaries and professors' salaries significant?

Yes, the difference between associate professors' salaries and professors' salaries is significant. The violin plot shows that professors tend to have higher salaries than associate professors.

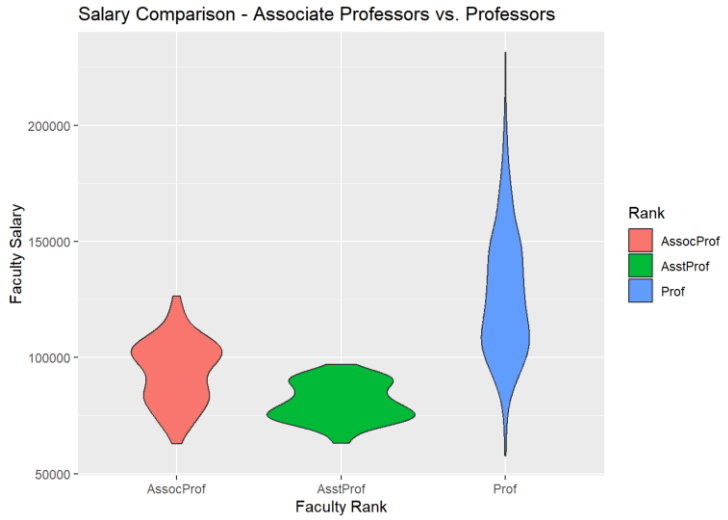

Figure 22. Salary Comparison - Associate Professors vs. Professors using violin plot.

d. Are there any wage gaps for employees with the same experience and rank?

Yes, there are wage gaps for employees with the same experience and rank. The scatter plot clearly shows differing salary levels among individuals with identical experience and rank, indicating the presence of wage gaps. This suggests potential unjustified disparities or discrimination in salary distribution.

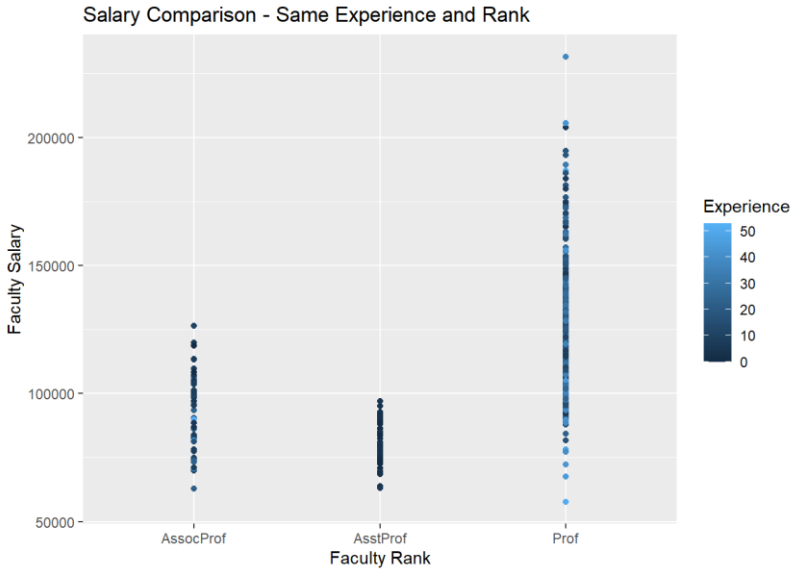

Figure 23. Salary Comparison - Same Experience and Rank using scatter plot.

9.0 Q7 If inequities exist, what are the suggested adjustment strategies that solve or improve the situation?

Inequities or wage gaps are identified within the dataset, several suggested adjustment strategies can help address or improve the situation:

1. Conduct a thorough analysis: Further investigate the factors contributing to the wage gaps or inequities, such as considering additional variables like gender, ethnicity, or department affiliation. This analysis can provide more insights into the underlying causes and help guide appropriate adjustment strategies.

2. Implement pay equity policies: Establish and enforce policies that ensure fair and equal compensation for employees with similar qualifications, experience, and responsibilities. These policies should be designed to address any wage gaps or discriminatory practices and promote pay equity within the organization.

3. Review and revise salary structures: Evaluate the current salary structures and consider adjustments that align with industry standards and best practices. This may involve revising salary scales, implementing performance-based pay systems, or addressing any discrepancies in pay levels based on rank or experience.

4. Provide professional development opportunities: Offer training, mentorship, and career advancement programs to employees to enhance their skills, knowledge, and qualifications. This can help create a more equitable environment by enabling employees to progress in their careers and increase their earning potential based on merit and achievement.

5. Foster transparency and communication: Ensure transparency in the compensation process by clearly communicating salary structures, criteria for promotions, and other relevant information to all employees. Encourage open dialogue and feedback mechanisms to address any concerns or perceptions of inequity and provide opportunities for employees to voice their opinions.

6. Regularly monitor and evaluate: Continuously monitor salary data, analyze trends, and conduct periodic reviews to identify any emerging inequities or wage gaps. Regular evaluations will help assess the effectiveness of adjustment strategies and allow for timely interventions when necessary.

It is important to note that specific adjustment strategies will depend on the organization's policies, legal requirements, and the unique characteristics of the workforce. Consulting with HR professionals, legal experts, or relevant stakeholders can provide valuable guidance in developing and implementing appropriate strategies to address wage gaps and promote equity. By implementing these adjustment strategies and continuously striving for fair and equitable compensation practices, organizations can work towards eliminating inequities and fostering a more inclusive and equitable work environment.

10.0 Conclusion:

our analysis of the faculty dataset from Datum College has yielded valuable insights into faculty composition and remuneration. Through thorough data cleansing and quality assessment, we have addressed various factors such as uniqueness, completeness, validity, consistency, relevancy, and timeliness. By eliminating duplicates, handling missing values, managing outliers, and rectifying misspelled ranks, we have significantly enhanced the dataset's overall quality and integrity.

The insights provided by this dataset can be instrumental in assessing Datum College's organizational structure, aiding in decision-making processes regarding faculty recruitment and compensation. Leveraging these valuable insights, Datum College can strive for continuous improvement and enhance the effectiveness of its faculty composition and remuneration strategies.

11.0 References

[1] Collibra, "The 6 dimensions of data quality," Collibra Blog, 2021. [Online]. Available: https://www.collibra.com/us/en/blog/the-6-dimensions-of-data-quality [Accessed May 25, 2023].
[2] GeeksforGeeks, "Data Cleaning in R," GeeksforGeeks, 2021. [Online]. Available: https://www.geeksforgeeks.org/data-cleaning-in-r/ [Accessed May 25, 2023].
[3] J. Rehmann, "Data Cleaning in R," Statistics Globe, 2021. [Online]. Available: https://statisticsglobe.com/data-cleaning-r/ [Accessed May 25, 2023].
[4] Precisely, "The 6 Dimensions of Data Quality: How to Measure Data Quality," Precisely Blog, 2021. [Online]. Available: https://www.precisely.com/blog/data-quality/data-quality-dimensions-measure [Accessed May 25, 2023].
[5] L. Zhang, "How to Explore a New Data Set," in L. Zhang's R Tips Book, 2021. [Online]. Available: https://bookdown.org/lyzhang10/lzhang_r_tips_book/how-to-explore-a-new-data-set.html [Accessed May 25, 2023].