# dfsdf

sdf

dsf

# sd

dfs

df

sdf

# Data Mining Project Documentation

# 1. Project Overview

**Title:** Data Mining on Ask A Manager Salary Survey 2021**Dataset Source:** Ask A Manager Salary Survey 2021 (Responses).xlsx (downloaded from the Ask A Manager website)**Objective:**

- Prepare and clean the dataset for further data mining analysis.
- Ensure that the dataset meets the project requirements (at least 10 columns with 2066 records).
- Apply various data mining algorithms for pattern discovery and predictive analysis.

**Dataset Details:**

- **Number of Columns:** 16 (project requirement was a minimum of 10 columns)
- **Number of Records:** 2066

---

# 2. Data Acquisition

The dataset was downloaded from the Ask A Manager website. The file, named **"Ask A Manager Salary Survey 2021 (Responses).xlsx"**, contains salary survey responses from professionals, providing a rich mix of categorical and numerical data that is suitable for various data mining techniques.

---

# 3. Data Cleaning and Preprocessing

Before applying any data mining algorithms, extensive data cleaning and transformation were performed using **Microsoft Excel** and **Kutools**. The following steps were taken:

## 3.1. Removing Unnecessary Columns

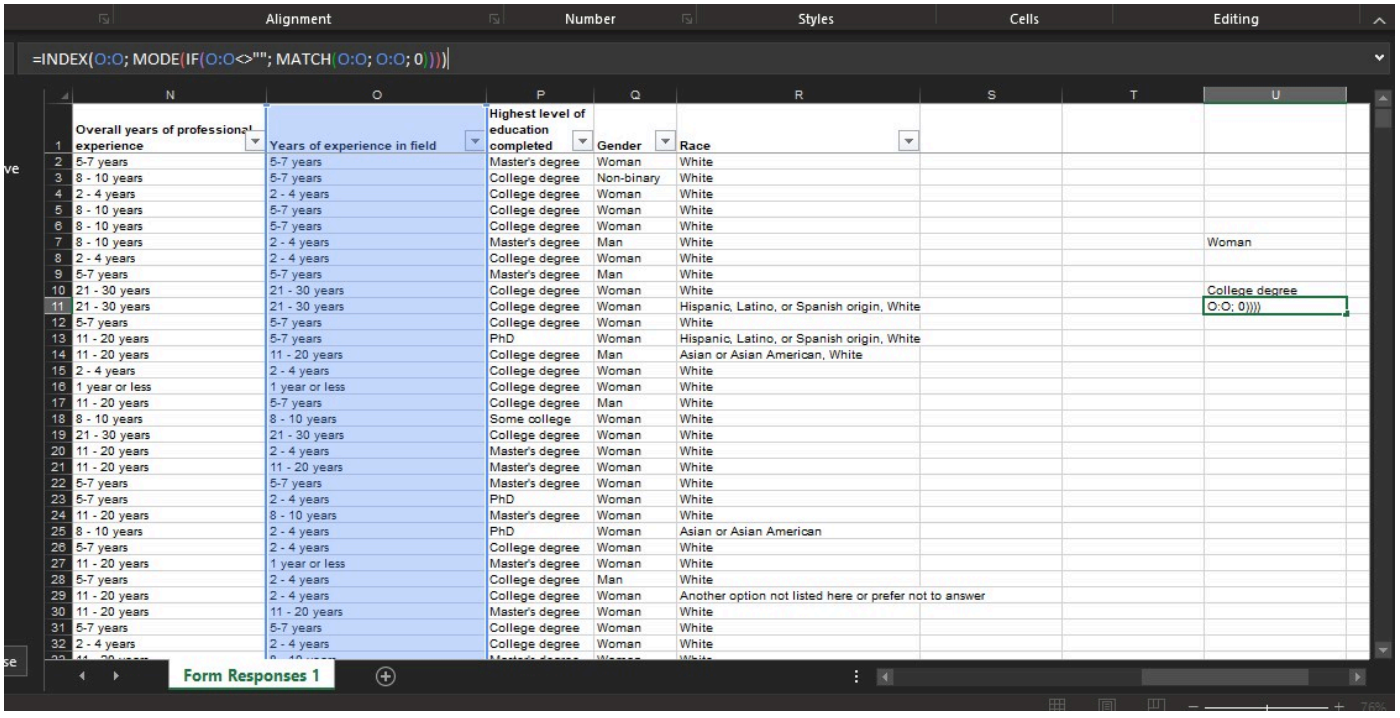
- **Rationale:** The original dataset contained **16** columns. Irrelevant or redundant columns were removed to focus on essential attributes, ensuring the dataset met the minimum requirement of **10 columns**.
- **Outcome:** A streamlined dataset that retains only the necessary columns.

# 3.2. Handling Missing Values

### 3.2.1. Missing Values in Categorical Data

- **Method:** Missing values in categorical fields (e.g., Gender) were imputed using the **mode**.
- **Example Formula (Excel):**

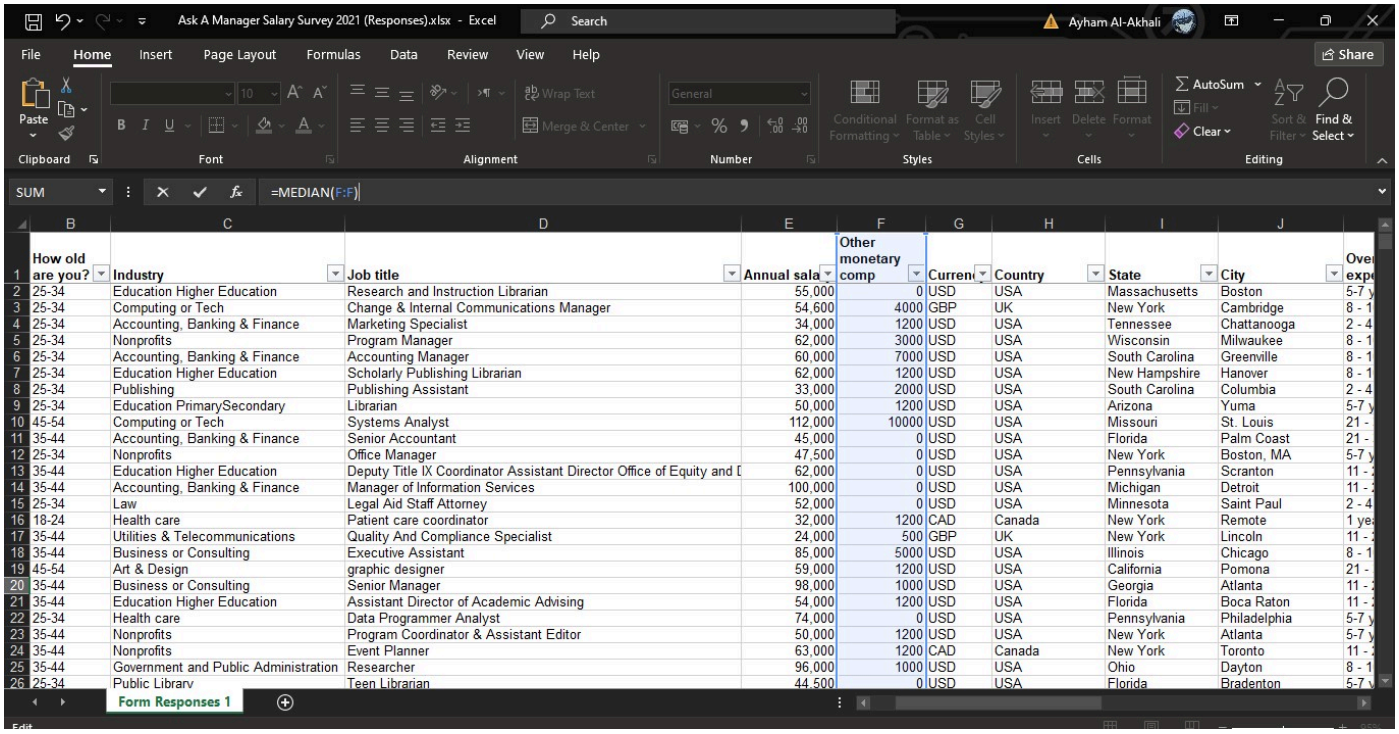
```
=INDEX(Q:Q; MODE(IF(Q:Q<>""; MATCH(Q:Q; Q:Q; 0))))
```




| | N | O | P | Q | R | S | T | U |
|---|---|---|---|---|---|---|---|---|
| 1 | Overall years of professional experience | Years of experience in field | Highest level of education completed | Gender | Race | | | |
| 2 | 5-7 years | 5-7 years | Master's degree | Woman | White | | | |
| 3 | 8 - 10 years | 5-7 years | College degree | Non-binary | White | | | |
| 4 | 2 - 4 years | 2 - 4 years | College degree | Woman | White | | | |
| 5 | 8 - 10 years | 5-7 years | College degree | Woman | White | | | |
| 6 | 8 - 10 years | 5-7 years | College degree | Woman | White | | | |
| 7 | 8 - 10 years | 2 - 4 years | Master's degree | Man | White | | | Woman |
| 8 | 2 - 4 years | 2 - 4 years | College degree | Woman | White | | | |
| 9 | 5-7 years | 5-7 years | Master's degree | Man | White | | | |
| 10 | 21 - 30 years | 21 - 30 years | College degree | Woman | White | | | College degree |
| 11 | 21 - 30 years | 21 - 30 years | College degree | Woman | Hispanic, Latino, or Spanish origin, White | | | O:O; 0)))) |
| 12 | 5-7 years | 5-7 years | College degree | Woman | White | | | |
| 13 | 11 - 20 years | 5-7 years | PhD | Woman | Hispanic, Latino, or Spanish origin, White | | | |
| 14 | 11 - 20 years | 11 - 20 years | College degree | Man | Asian or Asian American, White | | | |
| 15 | 2 - 4 years | 2 - 4 years | College degree | Woman | White | | | |
| 16 | 1 year or less | 1 year or less | College degree | Woman | White | | | |
| 17 | 11 - 20 years | 5-7 years | College degree | Man | White | | | |
| 18 | 8 - 10 years | 8 - 10 years | Some college | Woman | White | | | |
| 19 | 21 - 30 years | 21 - 30 years | College degree | Woman | White | | | |
| 20 | 11 - 20 years | 2 - 4 years | Master's degree | Woman | White | | | |
| 21 | 11 - 20 years | 11 - 20 years | Master's degree | Woman | White | | | |
| 22 | 5-7 years | 5-7 years | Master's degree | Woman | White | | | |
| 23 | 5-7 years | 2 - 4 years | PhD | Woman | White | | | |
| 24 | 11 - 20 years | 8 - 10 years | Master's degree | Woman | White | | | |
| 25 | 8 - 10 years | 2 - 4 years | PhD | Woman | Asian or Asian American | | | |
| 26 | 5-7 years | 2 - 4 years | College degree | Woman | White | | | |
| 27 | 11 - 20 years | 1 year or less | Master's degree | Woman | White | | | |
| 28 | 5-7 years | 2 - 4 years | College degree | Man | White | | | |
| 29 | 11 - 20 years | 2 - 4 years | College degree | Woman | Another option not listed here or prefer not to answer | | | |
| 30 | 11 - 20 years | 11 - 20 years | Master's degree | Woman | White | | | |
| 31 | 5-7 years | 5-7 years | College degree | Woman | White | | | |
| 32 | 2 - 4 years | 2 - 4 years | College degree | Woman | White | | | |

### 3.2.2. Missing Values in Numerical Data (Salary)

**Method:** Missing numerical values (e.g., Salary) were imputed using the **median** to reduce the influence of outliers.



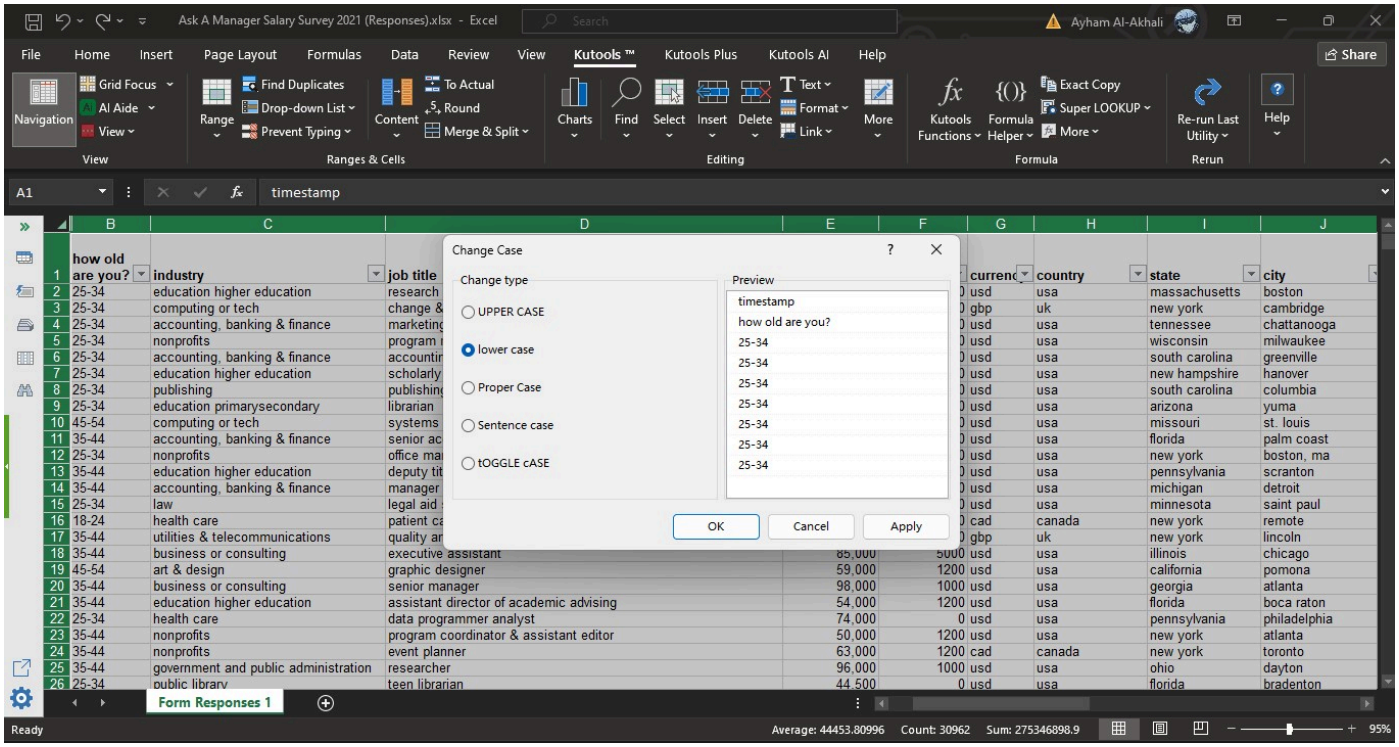

# 3.3. Standardizing Text Data

### Text Transformation:

All text entries were converted to lowercase for uniformity.

### Tool Used:

Kutools in Excel was used for bulk text conversion.



## Job Title Correction

**Issue:**

The dataset contained job titles with abbreviations like **"sr."**

**Solution:**

Replaced all instances of **"sr."** with **"senior"** to maintain consistency.

---

# 4. Data Mining Algorithms Applied

After cleaning the dataset, several data mining algorithms were applied to extract meaningful patterns and build predictive models. These algorithms include:

## 4.1. Apriori Algorithm

**Purpose:**

- Used for mining frequent itemsets and discovering association rules in the dataset.

**Application:**

- The Apriori algorithm helped in identifying relationships and co-occurrence patterns among categorical features, such as job roles, industries, and other survey responses.

---

## 4.2. Naive Bayes Classification

**Purpose:**

- Applied for predictive classification tasks using a probabilistic approach based on Bayes' theorem.

**Application:**

- Naive Bayes was used to predict categorical outcomes (for example, predicting the likelihood of a respondent belonging to a certain salary bracket or job category based on their survey responses).

---

# 5. Results and Next Steps

## 5.1. Outcomes

**Data Quality:**

- The dataset has been successfully cleaned, standardized, and preprocessed, ensuring high data quality and consistency.

**Algorithm Application:**

- **Apriori**: Revealed key association rules among survey responses.
- **Naive Bayes**: Provided a probabilistic model for class prediction, aiding in understanding categorical distributions.
- **ID3 Decision Tree**: Offered insight into the most significant attributes affecting the target variable.
- **K-Means Clustering**: Helped identify distinct segments within the dataset for further targeted analysis.

---

# 6. Conclusion

This project successfully transformed the raw **"Ask A Manager Salary Survey 2021"** dataset into a clean and standardized format ready for data mining. The process included:

**Data Cleaning:**

- Removing unnecessary columns, handling missing values, and standardizing text.

**Algorithm Application:**

- Utilizing the **Apriori**, **Naive Bayes**, **ID3 Decision Tree**, and **K-Means** algorithms to extract patterns, build predictive models, and segment the data.

The application of these algorithms has provided a multifaceted view of the dataset, setting the stage for further exploration and deeper analysis. Future work will focus on refining these models and interpreting the results to derive actionable business insights.