

- [dfsdf](#)
- [sd](#)
- [Data Mining Project Documentation](#)
  - [1. Project Overview](#)
  - [2. Data Acquisition](#)
  - [3. Data Cleaning and Preprocessing](#)
    - [3.1. Removing Unnecessary Columns](#)
    - [3.2. Handling Missing Values](#)
      - [3.2.1. Missing Values in Categorical Data](#)
      - [3.2.2. Missing Values in Numerical Data \(Salary\)](#)
    - [3.3. Standardizing Text Data](#)
      - [Text Transformation:](#)
      - [Tool Used:](#)
    - [Job Title Correction](#)
      - [Issue:](#)
      - [Solution:](#)
  - [4. Data Mining Algorithms Applied](#)
    - [4.1. Apriori Algorithm](#)
    - [4.2. Naive Bayes Classification](#)
  - [5. Results and Next Steps](#)
    - [5.1. Outcomes](#)
  - [6. Conclusion](#)

## 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))))
```

The screenshot shows an Excel spreadsheet with the following columns: Overall years of professional experience, Years of experience in field, Highest level of education completed, Gender, and Race. The data is organized into rows, with the first row being the header. The formula bar at the top displays the Excel formula: `=INDEX(O:O; MODE(IF(O:O<>""; MATCH(O:O; O:O; 0))))`. The spreadsheet is titled "Form Responses 1".

	N	O	P	Q	R	S	T	U
	Overall years of professional experience	Years of experience in field	Highest level of education completed	Gender	Race			
1	5-7 years	5-7 years	Master's degree	Woman	White			
2	8 - 10 years	5-7 years	College degree	Non-binary	White			
3	2 - 4 years	2 - 4 years	College degree	Woman	White			
4	8 - 10 years	5-7 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			
11	21 - 30 years	21 - 30 years	College degree	Woman	Hispanic, Latino, or Spanish origin, White			College degree
12	5-7 years	5-7 years	College degree	Woman	White			O:O; 0))
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			
33	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.

	B	C	D	E	F	G	H	I	J
	How old are you?	Industry	Job title	Annual salary	Other monetary comp	Current	Country	State	City
1	25-34	Education Higher Education	Research and Instruction Librarian	55,000	0	USD	USA	Massachusetts	Boston
2	25-34	Computing or Tech	Change & Internal Communications Manager	54,600	4000	GBP	UK	New York	Cambridge
3	25-34	Accounting, Banking & Finance	Marketing Specialist	34,000	1200	USD	USA	Tennessee	Chattanooga
4	25-34	Nonprofits	Program Manager	62,000	3000	USD	USA	Wisconsin	Milwaukee
5	25-34	Accounting, Banking & Finance	Accounting Manager	60,000	7000	USD	USA	South Carolina	Greenville
6	25-34	Education Higher Education	Scholarly Publishing Librarian	62,000	1200	USD	USA	New Hampshire	Hanover
7	25-34	Publishing	Publishing Assistant	33,000	2000	USD	USA	South Carolina	Columbia
8	25-34	Education Primary/Secondary	Librarian	50,000	1200	USD	USA	Arizona	Yuma
9	45-54	Computing or Tech	Systems Analyst	112,000	10000	USD	USA	Missouri	St. Louis
10	35-44	Accounting, Banking & Finance	Senior Accountant	45,000	0	USD	USA	Florida	Palm Coast
11	25-34	Nonprofits	Office Manager	47,500	0	USD	USA	New York	Boston, MA
12	35-44	Education Higher Education	Deputy Title IX Coordinator Assistant Director Office of Equity and	62,000	0	USD	USA	Pennsylvania	Scranton
13	35-44	Accounting, Banking & Finance	Manager of Information Services	100,000	0	USD	USA	Michigan	Detroit
14	25-34	Law	Legal Aid Staff Attorney	52,000	0	USD	USA	Minnesota	Saint Paul
15	18-24	Health care	Patient care coordinator	32,000	1200	CAD	Canada	New York	Remote
16	35-44	Utilities & Telecommunications	Quality And Compliance Specialist	24,000	500	GBP	UK	New York	Lincoln
17	35-44	Business or Consulting	Executive Assistant	85,000	5000	USD	USA	Illinois	Chicago
18	45-54	Art & Design	graphic designer	59,000	1200	USD	USA	California	Pomona
19	35-44	Business or Consulting	Senior Manager	98,000	1000	USD	USA	Georgia	Atlanta
20	35-44	Education Higher Education	Assistant Director of Academic Advising	54,000	1200	USD	USA	Florida	Boca Raton
21	25-34	Health care	Data Programmer Analyst	74,000	0	USD	USA	Pennsylvania	Philadelphia
22	35-44	Nonprofits	Program Coordinator & Assistant Editor	50,000	1200	USD	USA	New York	Atlanta
23	35-44	Nonprofits	Event Planner	63,000	1200	CAD	Canada	New York	Toronto
24	35-44	Government and Public Administration	Researcher	96,000	1000	USD	USA	Ohio	Dayton
25	25-34	Public Library	Teen Librarian	44,500	0	USD	USA	Florida	Bradenton

## 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.

FileHomeInsertPage LayoutFormulasDataReviewViewKutools™Kutools PlusKutools AIHelp

Navigation

Grid FocusAl AideView

RangeFind DuplicatesDrop-down ListPrevent Typing

View

Ranges & CellsContentTo ActualRoundMerge & Split

ChartsFindSelectInsertDeleteFormatLinkMore

Editing

Kutools FunctionsFormulaExact CopySuper LOOKUPMore

Formula

Re-run Last UtilityRun

Help

Share

A1

timestamp

	B	C	D	E	F	G	H	I	J
1	how old are you?	industry	job title			current	country	state	city
2	25-34	education higher education	research	timestamp		0	usa	massachusetts	boston
3	25-34	computing or tech	change & internal communications manager	timestamp		4000	gbp	uk	new york
4	25-34	accounting, banking & finance	marketing specialist	timestamp		1200	usa	tennessee	chattanooga
5	25-34	nonprofits	program manager	timestamp		3000	usa	wisconsin	milwaukee
6	25-34	accounting, banking & finance	accounting manager	timestamp		7000	usa	south carolina	greenville
7	25-34	education higher education	scholarly publishing librarian	timestamp		1200	usa	new hampshire	hanover
8	25-34	publishing	publishing assistant	timestamp		2000	usa	south carolina	columbia
9	25-34	education primary/secondary	librarian	timestamp		1200	usa	arizona	yuma
10	45-54	computing or tech	systems analyst	timestamp		10000	usa	missouri	st. louis
11	35-44	accounting, banking & finance	senior accountant	timestamp		0	usa	florida	palm coast
12	25-34	nonprofits	office manager	timestamp		0	usa	new york	boston, ma
13	35-44	education higher education	deputy title ix coordinator assistant director office of equity and	timestamp		0	usa	pennsylvania	scranton
14	35-44	accounting, banking & finance	manager of information services	timestamp		0	usa	michigan	detroit
15	25-34	law	legal aid staff attorney	timestamp		0	usa	minnesota	saint paul
16	18-24	health care	patient care coordinator	timestamp		1200	cad	canada	remote
17	35-44	utilities & telecommunications	quality and compliance specialist	timestamp		500	gbp	uk	new york
18	35-44	business or consulting	executive assistant	timestamp		5000	usa	illinois	chicago
19	45-54	art & design	graphic designer	timestamp		1200	usa	california	pomona
20	35-44	business or consulting	senior manager	timestamp		1000	usa	georgia	atlanta
21	35-44	education higher education	assistant director of academic advising	timestamp		1200	usa	florida	boca raton
22	25-34	health care	data programmer analyst	timestamp		0	usa	pennsylvania	philadelphia
23	35-44	nonprofits	program coordinator & assistant editor	timestamp		1200	usa	new york	atlanta
24	35-44	nonprofits	event planner	timestamp		1200	cad	canada	new york
25	35-44	government and public administration	researcher	timestamp		1000	usa	ohio	dayton
26	25-34	public library	teen librarian	timestamp		0	usa	florida	bradenton

Change Case

Change type

☐ UPPER CASE

☒ lower case

☐ Proper Case

☐ Sentence case

☐ TOGGLE CASE

Preview

timestamp

how old are you?

25-34

25-34

25-34

25-34

25-34

25-34

OK

Cancel

Apply

Form Responses 1

Ready

Average: 44453.80996Count: 30962Sum: 275346898.9

95%

# 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.