Data Mining Project

Master in Data Science and Advanced Analytics

**NOVA Information Management School**

Universidade Nova de Lisboa

Data Mining Project Report Part 1

**Group 50**

Joshua Wehr, 20240501

Benedikt Ruggaber, 20240500

Steven Burgerman, 69696969

Fall/Spring Semester 2024-2025

TABLE OF CONTENTS

[1. Introduction 1](#_Toc180427240)

[2. Data Exploration and Summary 1](#_Toc180427241)

[2.1. Data Exploration 1](#_Toc180427242)

[2.1.1. Level 3 title 2](#_Toc180427243)

[2.1.1.1. Level 4 title 2](#_Toc180427244)

[3. Another Section Heading 2](#_Toc180427245)

[Bibliographical References (Optional, Not included in page limit) 4](#_Toc180427246)

[Appendix A (Optional, Not included in page limit) 5](#_Toc180427247)

[Annexes (Optional, Not included in page limit) 6](#_Toc180427248)

# Introduction

The food delivery industry has experienced rapid growth and transformation in the era of big data, fueled by the vast amounts of data it generates. This report explores the application of data mining techniques to analyze customer behavior in the food delivery sector, using a comprehensive dataset from ABCDEats, a prominent food delivery service.

The ABCDEats dataset provides a rich source of customer information, including demographics, order history, cuisine preferences, temporal patterns, and promotional interactions. By leveraging this data, we aim to uncover meaningful patterns and insights that can inform business strategies and enhance customer experiences in the competitive food delivery market.

Our study focuses on several key aspects of customer behavior, including:

1. ⁠The relationship between customer demographics and cuisine preferences
2. ⁠Temporal patterns in ordering behavior
3. ⁠The impact of promotions on customer engagement
4. ⁠Customer segmentation based on ordering patterns and preferences

Through the application of various data mining techniques, including exploratory data analysis, clustering algorithms, and predictive modeling, we seek to extract actionable insights from this complex dataset. The findings of this study have potential implications for targeted marketing strategies, menu optimization, and overall service improvement in the food delivery industry.

This Report is Part 1 of three Parts to be delivery in scope of this Data Mining Project.

Part 1 focuses on Exploratory Data Analyses; the expected outcome of Part 1 of the Project is:

1. **Data Exploration and Summary:** Examine the dataset thoroughly, highlighting key statistics such as averages, distributions, and outliers for each feature. Discuss how these characteristics might influence the analysis.
2. **Trend and Pattern Identification:** Identify any notable trends, patterns, or anomalies within the data. Investigate relationships and correlations between different features.
3. **Feature Engineering:** Create additional features that could potentially improve the analysis, based on insights gained from the data exploration.
4. **Visualization for Insights:** Utilize visualizations to clearly present findings and support your analysis, enabling easier understanding of the data and the relationships between variables.

While this Report documents everything that has been done by the listed member during the project.

# Data Exploration and Summary

## Data Exploration

A few findings:

* We have 57 columns
* HR\_0 is Object, all other HR\_n are int64#
* Null Values for each column:
  + customer\_age: 727
  + first\_order: 106
* dtypes:
  + float64: 17
  + int64: 35
  + object: 4

## Data Inconsistencies

We began our analysis by evaluating possible inconsistencies in the data, including checking for duplicates, missing values, and possible incorrect data types. First, we counted the total number of duplicates in the data set. We found there to be 13 duplicates in the data and dropped these rows. Next, we corrected the data types for 3 variables: customer\_age, first\_order, and HR\_0. Each of these were floats in the original dataset, yet their values are represented by whole numbers rather than decimals, and so we converted them into integers. We made the decision to leave is\_chain as an integer even though the metadata implies a Boolean value. Upon close analysis, it appears that is\_chain represents the total number of chain restaurant orders made by the customer over the 3 month period rather than simply whether or not they ordered from any chain restaurants, and we treated it like such in the rest of our analysis.

## Statistical findings (tables)

Statistical Analyses

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Feature name | count | mean | std | min | 25% | 50% | 75% | max |
| vendor\_count | 31888.0 | 3.102609 | 2.771587 | 0.0 | 1.0 | 2.0 | 4.00 | 41.00 |
| product\_count | 31888.0 | 5.668245 | 6.957287 | 0.0 | 2.0 | 3.0 | 7.00 | 269.00 |
| is\_chain | 31888.0 | 2.818866 | 3.977529 | 0.0 | 1.0 | 2.0 | 3.00 | 83.00 |
| last\_order | 31888.0 | 63.675521 | 23.226123 | 0.0 | 49.0 | 70.0 | 83.00 | 90.00 |
| CUI\_American | 31888.0 | 4.880438 | 11.654018 | 0.0 | 0.0 | 0.0 | 5.66 | 280.21 |
| CUI\_Asian | 31888.0 | 9.960451 | 23.564351 | 0.0 | 0.0 | 0.0 | 11.83 | 896.71 |
| CUI\_Beverages | 31888.0 | 2.300633 | 8.479734 | 0.0 | 0.0 | 0.0 | 0.00 | 229.22 |
| CUI\_Cafe | 31888.0 | 0.801163 | 6.427132 | 0.0 | 0.0 | 0.0 | 0.00 | 326.10 |
| CUI\_Chicken Dishes | 31888.0 | 0.768096 | 3.657273 | 0.0 | 0.0 | 0.0 | 0.00 | 219.66 |
| CUI\_Chinese | 31888.0 | 1.431218 | 8.191755 | 0.0 | 0.0 | 0.0 | 0.00 | 739.73 |
| CUI\_Desserts | 31888.0 | 0.884359 | 5.259868 | 0.0 | 0.0 | 0.0 | 0.00 | 230.07 |
| CUI\_Healthy | 31888.0 | 0.950203 | 5.830590 | 0.0 | 0.0 | 0.0 | 0.00 | 255.81 |
| CUI\_Indian | 31888.0 | 1.631153 | 7.443234 | 0.0 | 0.0 | 0.0 | 0.00 | 309.07 |
| CUI\_Italian | 31888.0 | 3.233411 | 11.247990 | 0.0 | 0.0 | 0.0 | 0.00 | 468.33 |
| CUI\_Japanese | 31888.0 | 2.995379 | 10.180851 | 0.0 | 0.0 | 0.0 | 0.00 | 706.14 |
| CUI\_Noodle Dishes | 31888.0 | 0.711676 | 4.536457 | 0.0 | 0.0 | 0.0 | 0.00 | 275.11 |
| CUI\_OTHER | 31888.0 | 2.999913 | 9.768300 | 0.0 | 0.0 | 0.0 | 0.00 | 366.08 |
| CUI\_Street Food / Snacks | 31888.0 | 3.913253 | 15.548507 | 0.0 | 0.0 | 0.0 | 0.00 | 454.45 |
| CUI\_Thai | 31888.0 | 0.841697 | 4.433047 | 0.0 | 0.0 | 0.0 | 0.00 | 136.38 |
| DOW\_0 | 31888.0 | 0.555914 | 1.013601 | 0.0 | 0.0 | 0.0 | 1.00 | 16.00 |
| DOW\_1 | 31888.0 | 0.567486 | 1.044090 | 0.0 | 0.0 | 0.0 | 1.00 | 17.00 |
| DOW\_2 | 31888.0 | 0.591006 | 1.045907 | 0.0 | 0.0 | 0.0 | 1.00 | 15.00 |
| DOW\_3 | 31888.0 | 0.619449 | 1.069672 | 0.0 | 0.0 | 0.0 | 1.00 | 17.00 |
| DOW\_4 | 31888.0 | 0.677747 | 1.088122 | 0.0 | 0.0 | 0.0 | 1.00 | 16.00 |
| DOW\_5 | 31888.0 | 0.652973 | 1.069947 | 0.0 | 0.0 | 0.0 | 1.00 | 20.00 |
| DOW\_6 | 31888.0 | 0.704246 | 1.167446 | 0.0 | 0.0 | 0.0 | 1.00 | 20.00 |
| HR\_1 | 31888.0 | 0.053845 | 0.317013 | 0.0 | 0.0 | 0.0 | 0.00 | 14.00 |
| HR\_2 | 31888.0 | 0.063190 | 0.351498 | 0.0 | 0.0 | 0.0 | 0.00 | 12.00 |
| HR\_3 | 31888.0 | 0.118759 | 0.500862 | 0.0 | 0.0 | 0.0 | 0.00 | 11.00 |
| HR\_4 | 31888.0 | 0.101700 | 0.437493 | 0.0 | 0.0 | 0.0 | 0.00 | 14.00 |
| HR\_5 | 31888.0 | 0.081943 | 0.358705 | 0.0 | 0.0 | 0.0 | 0.00 | 7.00 |
| HR\_6 | 31888.0 | 0.069681 | 0.329461 | 0.0 | 0.0 | 0.0 | 0.00 | 8.00 |
| HR\_7 | 31888.0 | 0.076800 | 0.377700 | 0.0 | 0.0 | 0.0 | 0.00 | 15.00 |
| HR\_8 | 31888.0 | 0.131899 | 0.635582 | 0.0 | 0.0 | 0.0 | 0.00 | 52.00 |
| HR\_9 | 31888.0 | 0.233912 | 0.724906 | 0.0 | 0.0 | 0.0 | 0.00 | 23.00 |
| HR\_10 | 31888.0 | 0.329560 | 0.891161 | 0.0 | 0.0 | 0.0 | 0.00 | 25.00 |
| HR\_11 | 31888.0 | 0.378167 | 0.959961 | 0.0 | 0.0 | 0.0 | 0.00 | 36.00 |
| HR\_12 | 31888.0 | 0.314162 | 0.842484 | 0.0 | 0.0 | 0.0 | 0.00 | 26.00 |
| HR\_13 | 31888.0 | 0.236453 | 0.637502 | 0.0 | 0.0 | 0.0 | 0.00 | 14.00 |
| HR\_14 | 31888.0 | 0.215630 | 0.599006 | 0.0 | 0.0 | 0.0 | 0.00 | 13.00 |
| HR\_15 | 31888.0 | 0.277032 | 0.738162 | 0.0 | 0.0 | 0.0 | 0.00 | 23.00 |
| HR\_16 | 31888.0 | 0.356435 | 0.874449 | 0.0 | 0.0 | 0.0 | 0.00 | 22.00 |
| HR\_17 | 31888.0 | 0.390962 | 0.943721 | 0.0 | 0.0 | 0.0 | 0.00 | 20.00 |
| HR\_18 | 31888.0 | 0.336961 | 0.893949 | 0.0 | 0.0 | 0.0 | 0.00 | 24.00 |
| HR\_19 | 31888.0 | 0.245610 | 0.795296 | 0.0 | 0.0 | 0.0 | 0.00 | 35.00 |
| HR\_20 | 31888.0 | 0.142812 | 0.586529 | 0.0 | 0.0 | 0.0 | 0.00 | 36.00 |
| HR\_21 | 31888.0 | 0.071155 | 0.348536 | 0.0 | 0.0 | 0.0 | 0.00 | 11.00 |
| HR\_22 | 31888.0 | 0.048263 | 0.298265 | 0.0 | 0.0 | 0.0 | 0.00 | 15.00 |
| HR\_23 | 31888.0 | 0.045189 | 0.282006 | 0.0 | 0.0 | 0.0 | 0.00 | 13.00 |

Outlier Analysis

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Feature | Lower Bound (Outliers) | Upper Bound (Outliers) | Max | Outliers Detected |
| vendor\_count | -3.5 | 8.5 | 41.0 | Yes, values above 8.5 |
| product\_count | -7.5 | 16.5 | 269.0 | Yes, values above 16.5 |
| is\_chain | -3.0 | 7.0 | 83.0 | Yes, values above 7 |
| last\_order | 16.0 | 116.0 | 90.0 | No |
| CUI\_American | -8.49 | 14.15 | 280.21 | Yes, values above 14.15 |
| CUI\_Asian | -17.75 | 29.58 | 896.71 | Yes, values above 29.58 |
| CUI\_Beverages | -12.71 | 12.71 | 229.22 | Yes, values above 12.71 |
| CUI\_Cafe | -9.64 | 9.64 | 326.1 | Yes, values above 9.64 |
| CUI\_Chicken Dishes | -5.49 | 5.49 | 219.66 | Yes, values above 5.49 |
| CUI\_Chinese | -12.29 | 12.29 | 739.73 | Yes, values above 12.29 |
| CUI\_Desserts | -7.89 | 7.89 | 230.07 | Yes, values above 7.89 |
| CUI\_Healthy | -8.75 | 8.75 | 255.81 | Yes, values above 8.75 |
| CUI\_Indian | -11.16 | 11.16 | 309.07 | Yes, values above 11.16 |
| CUI\_Italian | -16.87 | 16.87 | 468.33 | Yes, values above 16.87 |
| CUI\_Japanese | -15.27 | 15.27 | 706.14 | Yes, values above 15.27 |
| CUI\_Noodle Dishes | -6.8 | 6.8 | 275.11 | Yes, values above 6.8 |
| CUI\_OTHER | -14.65 | 14.65 | 366.08 | Yes, values above 14.65 |
| CUI\_Street Food / Snacks | -23.32 | 23.32 | 454.45 | Yes, values above 23.32 |
| CUI\_Thai | -6.65 | 6.65 | 136.38 | Yes, values above 6.65 |
| DOW\_0 | -1.5 | 2.5 | 16.0 | Yes, values above 2.5 |
| DOW\_1 | -1.5 | 2.5 | 17.0 | Yes, values above 2.5 |
| DOW\_2 | -1.5 | 2.5 | 15.0 | Yes, values above 2.5 |
| DOW\_3 | -1.5 | 2.5 | 17.0 | Yes, values above 2.5 |
| DOW\_4 | -1.5 | 2.5 | 16.0 | Yes, values above 2.5 |
| DOW\_5 | -1.5 | 2.5 | 20.0 | Yes, values above 2.5 |
| DOW\_6 | -1.5 | 2.5 | 20.0 | Yes, values above 2.5 |
| HR\_1 - HR\_23 | Varies, generally < 0 | Varies, generally > 5 | Varies | Outliers found in higher HR values |

OLD:

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Feature | Count | Mean | Std Dev | Min | 25% | 50% (Median) | 75% | Max | Outliers |
| customer\_age | 31,161 | 27.51 | 7.16 | 15.0 | 23.0 | 26.0 | 31.0 | 80.0 | Max is high; outlier over 75 years |
| vendor\_count | 31,888 | 3.10 | 2.77 | 0.0 | 1.0 | 2.0 | 4.0 | 41.0 | Max is high; outlier over 15 vendors |
| product\_count | 31,888 | 5.67 | 6.96 | 0.0 | 2.0 | 3.0 | 7.0 | 269.0 | Max is an extreme outlier; any over 50 considered |
| is\_chain | 31,888 | 2.82 | 3.98 | 0.0 | 1.0 | 2.0 | 3.0 | 83.0 | Max is an extreme outlier; any over 15 considered |
| first\_order | 31,782 | 28.48 | 24.11 | 0.0 | 7.0 | 22.0 | 45.0 | 90.0 | No significant outliers |
| last\_order | 31,888 | 63.68 | 23.23 | 0.0 | 49.0 | 70.0 | 83.0 | 90.0 | No significant outliers |
| CUI\_American | 31,888 | 4.88 | 11.65 | 0.0 | 0.0 | 0.0 | 5.66 | 280.21 | Max is an extreme outlier; over 50 is significant |
| CUI\_Asian | 31,888 | 9.96 | 23.56 | 0.0 | 0.0 | 0.0 | 11.83 | 896.71 | Max is an extreme outlier; any over 100 |
| CUI\_Beverages | 31,888 | 2.30 | 8.48 | 0.0 | 0.0 | 0.0 | 0.0 | 229.22 | Max is an extreme outlier; any over 50 |
| CUI\_Cafe | 31,888 | 0.80 | 6.43 | 0.0 | 0.0 | 0.0 | 0.0 | 326.1 | Max is an extreme outlier; any over 50 |
| HR\_14 | 31,888 | 0.22 | 0.60 | 0.0 | 0.0 | 0.0 | 0.0 | 13.0 | Values over 5 are outliers |
| HR\_15 | 31,888 | 0.28 | 0.74 | 0.0 | 0.0 | 0.0 | 0.0 | 23.0 | Values over 10 are outliers |
| HR\_16 | 31,888 | 0.36 | 0.87 | 0.0 | 0.0 | 0.0 | 0.0 | 22.0 | Values over 10 are outliers |
| HR\_17 | 31,888 | 0.39 | 0.94 | 0.0 | 0.0 | 0.0 | 0.0 | 20.0 | Values over 10 are outliers |
| HR\_18 | 31,888 | 0.34 | 0.89 | 0.0 | 0.0 | 0.0 | 0.0 | 24.0 | Values over 10 are outliers |
| HR\_19 | 31,888 | 0.25 | 0.80 | 0.0 | 0.0 | 0.0 | 0.0 | 35.0 | Values over 10 are outliers |
| HR\_20 | 31,888 | 0.14 | 0.59 | 0.0 | 0.0 | 0.0 | 0.0 | 36.0 | Values over 10 are outliers |
| HR\_21 | 31,888 | 0.07 | 0.35 | 0.0 | 0.0 | 0.0 | 0.0 | 11.0 | Values over 5 are outliers |
| HR\_22 | 31,888 | 0.05 | 0.30 | 0.0 | 0.0 | 0.0 | 0.0 | 15.0 | Values over 5 are outliers |
| HR\_23 | 31,888 | 0.05 | 0.28 | 0.0 | 0.0 | 0.0 | 0.0 | 13.0 | Values over 5 are outliers |

## Histograms

A group of blue and white bars

Description automatically generated

* customer\_age:
  + The distribution appears right skewed, with a peak around the age range of 2030.
  + The mean age is 27.5 years old, which means the average customer is 27.5 years old. The data is right skewed, which indicates that while the majority of customers are similar to 27.5 years old, there are a few outlier customers which are significantly older. The right-skewness also indicates that using the median, 26.0 years old, is a better representation of the “average” customer than the mean of 27.5.
  + Most customers seem to be young adults, possibly indicating a target audience for a service or product appealing to younger demographics.
  + The number of customers decreases significantly as age increases, suggesting less engagement from older age groups.
* vendor\_count:
  + The distribution is also right skewed, with most vendors having fewer than 10 counts.
  + The median value is 2.0, which indicates that the average customer orders from 2 different vendors over 3 months. Median should be used over mean for skewed data (mean = 3.1).
  + There are a few instances with higher vendor counts (up to 40+), which may indicate outliers or vendors with a higher level of engagement.
* product\_count:
  + Most data points are clustered near the lower end, with many having fewer than 20 products.
  + Median product count is 3, meaning that the average customer orders 3 total products over the 3 months. (Mean = 5.67)
  + There are a few cases where the product count goes up to 250, indicating potential outliers or special cases with high product variety.
* is\_chain:
  + The median is 2.0, which means the average customer orders from 2 chain restaurants over 3 months. (Mean = 2.82)
  + However, there are some vendors with values reaching above 80, indicating chain establishments with potentially many locations or branches.
* first\_order:
  + The distribution is right skewed, indicating that the first order mostly occurs within the first few weeks.
  + The median of 22 indicates that the average customer ordered their first order 22 days from the start of the data set.
  + There are fewer customers who make their first order later, possibly due to declining interest or delayed engagement (I think this is likely because most customers placed their FIRST order earlier).
* last\_order:
  + Left-skewed. Meaning most customers likely placed their last order closer to the end of the data window, but a few outliers stopped ordering significantly earlier. Median is 70.0, which means that the average customer placed their last order 70 days after the start of the dataset.
* CUI\_American:
  + The histogram is heavily right skewed, with most values near zero.
  + This suggests that American cuisine orders are not very frequent, but there are a few cases with significantly higher orders, which may be outliers.
  + Median is 0.0, which means the average customer spends $0 on American food delivery over a 3 month period (Not me, a “Certified FATinum boy”).
* HR\_20:
  + Most values are clustered near zero, indicating that activity at this hour is infrequent.
  + However, there are outliers present, with some instances reaching up to 35 orders, which suggests some customers LOVE to order at 8pm.
  + Median is 0.0, which indicates the average customer orders nothing 8pm,
    - We might need to think a little more about this. The median for all HR is 0.0.

Overall Trends:

The data shows a tendency toward right skewed distributions across many variables, indicating that most values are concentrated at the lower end with some outliers at the higher end.

There are potential outliers in the data that may need further investigation, such as high product counts, vendor counts, and chain values. These outliers will likely be the targets of specific advertisements as they represent the most engagement.

The histograms suggest that most customers or orders are clustered around specific ranges, with decreasing frequency for higher values.

Figure 2.1 – Illustrative figure

Note that figure labels should be included after the figure. Sample text with the inclusion of figures and tables Sample text with the inclusion of figures and tables Sample text with the inclusion of figures and tables Sample text with the inclusion of figures and tables Sample text with the inclusion of figures and tables Sample text with the inclusion of figures and tables.

Table 2.1 – Illustrative table

|  |  |
| --- | --- |
| **Title** | **Title** |
| Text | Number |
| Text | Number |
| Text | Number |

The student can freely choose the table design, as long as it remains consistent throughout the document. Note that table labels should always be included before the table. Sample text with the inclusion of figures and tables Sample text with the inclusion of figures and tables Sample text with the inclusion of figures and tables Sample text with the inclusion of figures and tables Sample text with the inclusion of figures and tables Sample text with the inclusion of figures and tables.

### Level 3 title

Example of an unnumbered list: Notes from Steven to Steven

Put some feature engineering shit here with bar graphs and box plots dad

* Item 1
* Item 2
* Item 3

#### Level 4 title

Example of a numbered list:

1. Item 1
2. Item 2
3. Item 3

# Another Section Heading

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Duis eget arcu nunc. Suspendisse ac volutpat nisl, at ullamcorper purus. Vestibulum tempus vehicula mauris vitae imperdiet. Morbi nibh nisl, dignissim et tellus eget, volutpat rhoncus nisi. Maecenas eget ipsum massa. Quisque malesuada nulla a felis imperdiet, a placerat ipsum ornare.

Ut a neque eu nulla aliquam tincidunt sit amet a nisi. Nam vulputate, diam non pellentesque condimentum, erat nunc suscipit turpis, a consequat ipsum ex non sapien.

# Bibliographical References (Optional, Not included in page limit)

Use APA Style for the entire document

We suggest that students use a reference manager system (Zotero, Mendeley, EndNote),

Please review the style guide at: <https://apastyle.apa.org/style-grammar-guidelines/references/examples>:

Author, A. A., Author, B. B., & Author, C. C. (Year). Title of article. *Title of Periodical, volume number* (issue number), pages.

# Appendix A (Optional, Not included in page limit)

[Appendixes are for materials, tables, or more explanation material only done by the student]

# Annexes (Optional, Not included in page limit)

[Annexes are optional, since they have material and sources not developed by the students, so in most cases referencing them is enough]